கணினி மொழியியல்

முனைவர் ச. இராசேந்திரன்

பணிநிறை பேராசிரியர், மொழியியல் துறை

தமிழ்ப் பல்கலைக் கழகம்

இயல் 1:அறிமுகம்

செய்திகளை 1, 0 என்ற எண்களால் ஆய்வதுதான் கணிப்பொறி. மனித மூளை 0,1 என்ற அடிப்படை எண்களை வைத்துக் கொண்டு எந்திர மொழி மூலம் சிக்கலான கணிப்பொறியை உருவாக்கியது. எழுத்துக்களும் எண்களும் தனித்தன்மையான குறியீடுகளும் (Special Symbols) 0,1 என்பவைகளால் தான் உருப்படுத்தம் செய்யப்படுகின்றன. உயர்நிலை மொழிகள், பொருள்கோளிகள் (Inyterpreters) ஒன்றுகூட்டிகள் (Assemblers) என்பன மொழியியலால் பயன்பெற்றன. இப்பொழுது மொழியியலார் கணினியை அவர்களின் கடினமான செயல்பாடுகளைச் செய்யும் வன்மையான கருவியாக பயன்படுத்தும் காலம் வந்துவிட்டது.

இயற்கை மொழியும் செயற்கை மொழியும்

மனிதர்கள் இயற்கையாகப் பயன்படுத்தும் எந்த மொழியும் இயற்கை மொழியாகும். செயற்கையாக மனிதனால் உருவாக்கப்பட்ட மொழிகள் இயற்கை மொழியிலிருந்து வேறுபட்டவை. எடுத்துக்காட்டாக, கணிப்பொறியில் பயன்படுத்தும் Programing மொழிகளான போட்ரான் (Fortron), பேசிக் (Basic), சி புரோலாக் (Prolog), லிஸ்ப் (Lisp), பியால் (Perl) என்பனவெல்லாம் செயற்கை மொழிகளாகும்.

இயற்கை மொழியாய்வு (Natural Language Processing)

இயற்கை மொழியாய்வு கணிப்பொறியைக் கொண்டு மொழியை ஆய்வு செய்யும் எல்லா நடவடிக்கைகளையும் உள்ளடக்கும். பேச்சையும் (Speech) எழுத்து மொழியையும் ஆயும் கணினி நுட்பங்கள் இதில் அடங்கும்.

பேச்சு உருவாக்கம் (Speech Synthesis)

இயற்கை ஒலியான பேச்சை உருவாக்குவது நுட்ப அடிப்படையில் சிக்கலானதாகும். என்ன பேசப்படுகிறது என்பதைப் புரிந்துக்கொள்ள சரியான இசையோட்டம் போன்றவை தேவை.

பேச்சைத் தெரிந்துக் கொள்ளல் (Speech Recognition)

இது அடிப்படையில் தொடர்ச்சியான ஒலி அலைகளை வேறுபட்ட சொற்களாகப் பகுப்பது.

இயற்கை மொழியைப் புரிந்துக் கொள்ளுதல் (Natural Language Understanding)

தனித்தனி சொற்களில் இருந்து பொருண்மைக்குச் செல்லுதல்.

இயற்கை மொழியை உருவாக்குதல் (Natural Language Generatioin)

அனுமானிக்க இயலாத உள்ளீடுகளுக்குப் பொருத்தமான பதில்களை உருவாக்குதல்.

இயந்திர மொழிப்பெயர்ப்பு (Machine Translation)

ஒரு இயற்கை மொழியை இன்னொன்றாக மொழிப்பெயர்ப்பு செய்தல். பொதுவாக இயற்கை மொழியாய்வு பேச்சை உட்படுத்தினால் தான் பயன்படுத்தப்படுகிறது. பேச்சு மற்றும் இயற்கை மொழியாய்வு (Speech and Natural Processing NLP) என்பதை பேச்சையும் பிற இயற்கை மொழியாய்வுகளையும் உள்ளடக்கப் பயன்படுத்தலாம்.

சிக்கலான மொழிப் பழக்கத்தில் ஈடுபடத் தேவையான ஆறு வேறுபட்ட பாகுபாடுகளாகப் பிரிக்கலாம்.

1. **ஒலியனியல் (Phonology):**

மொழி ஒலிகளைப் பற்றி படிப்பது.

2. **உருபனியல் (Morphology)**:

சொற்களின் பொருண்மைக் கூறுகளைப் பற்றி படிப்பது.

3. **தொடரியல் (Syntex):**

சொற்களுக்கு இடையே உள்ள அமைப்பு உறவுகளைப்பற்றி படிப்பது.

4. **பொருண்மையியல் (Sementics)**

பொருண்மையைப் பற்றி படிப்பது.

5. **பயன் வழியியல்** (Pragmatics):

இலக்கை அடைய மொழி எவ்வாறு பயன்படுத்தப்படுகின்றது என்பதைப் பற்றி படிப்பது.

6. **கருத்தாடல் (Discorse)**

ஒரு கூற்றைவிட பெரிய மொழி அலகைப் பற்றி படிப்பது.

மொழியின் பல நிலைகளில் மொழிப்பகுப்பாய்வு (Language Processing at Different Levels of Language)

ஒலியியல்:

ஒலி அடையாள குறிகளிடம் இருந்து (Speech Signals) அல்லது கேட்பு அடையாளக் குறிகளிடமிருந்து (Audio Signals) சொற்களைப் பிரித்தெடுக்கும் பேச்சைத் தெரிந்து கொள்ளும் (Speech Recognition) செயல்பாடு.

ஒலிகள்--> சொற்கள்

/b/+/ɔ/+t-->/bɔt/ ‘boat’

க்+ஆ+க்+அ+ம்-->காகம்

உருபனியல்:

உருபன்களிலிருந்து சொற்களைப்பெறல். சொற்களை உருபன்களாகப்பகுத்தல்.

உருபன்கள்--> சொற்கள்

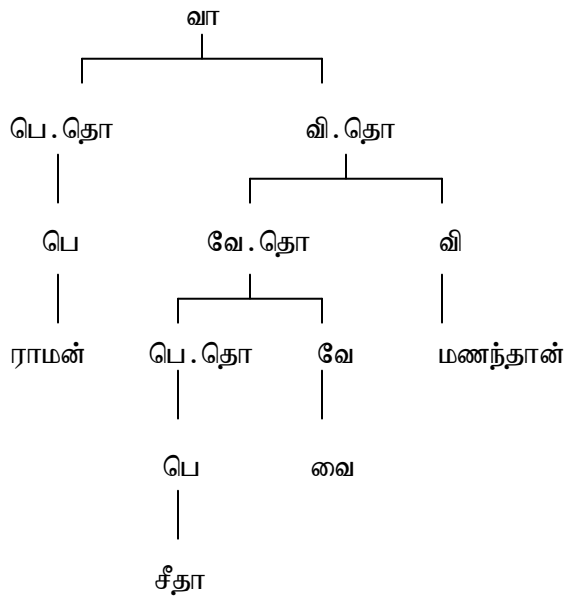
மரம் + கள் --> மரங்கள்

தொடரியல்:

சொற்களின் வரிசை அமைப்பிலிருந்து வாக்கிய அமைப்பைப் பெறல். வாக்கியங்கள் சொற்களாகப் பிரித்தல்.

சொல்வரிசை --> வாக்கிய அமைப்பு

‘ராமன்’ ‘சீதாவை’ மனந்தான் என்ற சொல் வரிசையிலிருந்து பின்வரும் பகுத்தாயப்பட்ட கிளை அமைப்பைப் பெறல்.



பொருண்மையியல்

வாக்கிய அமைப்பு, சொற்பொருண்மை இவற்றிலிருந்து வாக்கியப் பொருண்மையைப் பெறல்.

எடுத்துக்காட்டாக ராமன் சீதாவை மணந்தான் என்ற பகுத்தாயப்பட்ட கிளை அமைப்பிலிருந்து பின்வரும் தர்க்கக் கூற்றைப் பெறுவது.

மணந்தான் (ராமன், சீதா)

பயன்வழியியல்

வாக்கியப் பொருள் மற்றும் சூழல் இவற்றிலிருந்து சரியான பொருளைப் பெறல். இது பொது அறிவையும் முன்னர் வந்த கூற்றின் அறிவையும் உள்ளடக்கியது.

அவன் சீதாவை மணந்தான்

என்ற வாக்கியத்தில் வரும் பதிலீடுபெயர் (அவன்) முன்னர் வந்த வாக்கியத்தின் அறிவால் அவன் = ராமன் என்று புரிந்து கொள்ளப்படும்.

மொழி ஆய்வு மொழியைப் புரிந்து கொள்ளுதல் என்பதுடன் பின்வருவனவற்றையும் உள்ளடக்கும்.

1. பனுவல்கள் அல்லது பேச்சுச் செய்திகளைப் பொருள் கொண்டு புரிந்து கொள்ளுதல்
2. மொழி பெயர்த்தல்
3. பனுவல்கள் அல்லது பேச்சுச் செய்திகளை உருவாக்குதல்.

இயல் 2

சீரான வெளிப்பாடுகளும் தானியங்கிகளும்

(Regular expressions and Automata)

சீரான வெளிப்பாடுகள் (Regular expressions)

பனுவல் தொடர் வரிசைகளைப் (Text Sequnces) பண்பாக்கம் செய்யும் நிலை பெறுகுறிமானம் (Standard Notation) சீரான வெளிப்பாடாகும் (Regular expression). சீரான வெளிப்பாடு வலையில் தேடல் போன்ற சூழல்களில் பனுவல் கோர்வைகளை (Text Strings) குறிப்பிட்டுக் கூறவும் பிற செய்தி மீட்பு (Information Retrieval) பயன்பாடுகளிலும் பயன்படுத்தப்படுவதோடு சொல் ஆய்விலும், விரிதரவுகளிலிருந்து (Corpore) வருகை எண்ணிக்கைகளைக் கணக்கிடல் போன்ற செயல்பாடுகளிலும் முக்கியமான பங்கு வகிக்கின்றது.

மிக எளிதான சீரான வெளிப்பாட்டு வகை தனி எழுத்துக்களின் ஒரு தொடர்வரிசையாகும். எடுத்துக்காட்டாக, Wood Chuck என்பதைத் தேட நாம் /Wood chuck/ என்று தட்டச்சு செய்கின்றோம். எனவே /Butter cup/ என்ற சீரான வெளிப்பாடு Buttercup என்ற துணைக் கோர்வையை (Substring) உள்ளடக்கும் எந்த கோவையுடனும் பொருத்தம் பெறும்.

தேடல் கோர்வை (Search String)

ஒரு எழுத்தோ (எடுத்துக்காட்டாக/1/அல்லது ஒரு தொடர் வரிசையான எழுத்துக்களைக் கொண்டிருக்கலாம்.

எடுத்துக்காட்டாக /Urgl/.

சீரான வெளிப்பாடுகள்	பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு
/Wood chuks/	"Intersting links to wood chucks and lemurs"
/a/	"Mary Ann stoped by mona's"
/claire, sangs/	"Dogmor my gift please "claire says" "(all our prietiy songs)"
/sing/	"you we left the burgular behind again ! sad nore
/!/	

சீரான வெளிப்பாடுகள் எழுத்துக்கள் சின்ன வடிவ வகையைச் சார்ந்தனவா பெரிய வடிவ வகையைச் சார்ந்தனவா என்ற உணர்வு உள்ளன. எடுத்துக்காட்டாக சின்ன வடிவ வகையைச் சார்ந்த /s/ பெரிய வடிவ வகையைச் சார்ந்த /S/ -இல் இருந்து வேறுபடும். அதாவது /Wood chucks/ என்பது /Wood chucks/ என்பதுடன் பொருத்தமுறாது. செவ்வக அடைப்புக் குறிகளுக்குள் வரும் எழுத்துக்களின் கோர்வை எழுத்துக்களின் பிரிநிலையைக் (Disjunation) குறிப்பிடும். எடுத்துக்காட்டாக (Ww) என்ற அமைப்பொழுங்கு w அல்லது W உள்ள அமைப்பொழுங்குடன் பொருத்தமுறும்.

சீரான வெளிப்பாடுகள்	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/(Ww)ood chuck/	wood chuck –1	“wood chuck”
/(abc)/	wood chuck -2	“Near the pond”
/(123456789)/	எதாவது ஒரு எண்	planty of 2 to 5

சீரான வெளிப்பாடான (1234567890) என்பது எந்த ஒரு ஒற்றை எண்ணையும் குறிப்பிடும் எழுத்துக்களை இவ்வாறு குறிப்பிடுவது நன்றாக இராது.

(abcdefghijklmnopqrstuvwxyz)

‘_’ என்ற குறியைப் பயன்படுத்தலாம். எடுத்துக்காட்டாக /(b-g)/என்பது b,c,c,d,e,f,g, என்ற எதாவது எழுத்தைக் குறிப்பிடும். இதுப் போல /(2-5) / என்ற அமைப்பொழுங்கு 2,3,4,5 என்ற எதாவது ஒற்றை ஸ்தான எண்ணை குறிப்பிடும்.

சீரான வெளிப்பாடு	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
/(A-Z)/	பெரிய வடிவ வகை எழுத்து	“We met Ramasamy’
/(a-z)/	சின்ன வடிவ வகை எழுத்து	“My book”
/(O-9)/	ஒரு இலக்க எண்	“Chapter!

‘^’ (‘காரட் ‘Caret’) என்ற குறியீட்டின துணையுடன் செவ்வக அடைப்புக் குறியை எந்த ஒரு எழுத்து இல்லை என்பதைக் குறிப்பிடவும் பயன்படுத்தலாம். எடுத்துக்காட்டாக /(a)/ என்பது a- ஐத் தவிர வேறு எந்த எழுத்துடனும் பொருத்தமுறும் ‘^’ என்பது முதல் குறியீடாக இருக்கும் போது தான் இது உண்மையாகும்.

சீரான வெளிப்பாடு	**பொருத்தம்**	**பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு**
(A-Z)	பெரிய வடிவ வகை எழுத்து அல்ல	“Oyfn pripetchik”
(^Ss)	‘S’ – ஓ s-ஓ அல்ல	“I have no objection”
(^\.)	முற்றுப்புள்ளி அல்ல	“ Our house”
(e^)	e-ஓ ‘^’ ஓ	
a^b	‘a^b’ என்ற அமைப்பொழுங்கு	

எவ்வாறு நாம் Wood chuck என்பதையும் woodchucks என்பதையும் குறிப்பிட இயலும்? என்பது தற்போதைய பிரச்சினை. இதற்கு நாம் கேள்விக்குறியை /?/ பயன்படுத்தலாம்; கேள்விக்குறி முன்வரும் எழுத்தையோ ஒன்றும் இல்லாமையோ (பூஜியத்தையோ) குறிப்பிடும்.

சீரான வெளிப்பாடு	**பொருத்தம்**	**பொருந்தும் அமைப்பொகுங்குகளின் எடுத்துக்காட்டு**
Wood chucks	Wood chuck –ஓ Wood chucks -X	“Wood chuck”
Colou?r	Colour-ஓ Colour –ஓ	“Colour”

எத்தனை தடவை எழுத்து வர வேண்டும் என்பதை எவ்வாறு குறிப்பிடுவது என்பது நமது அடுத்த பிரச்சனை. சில வேளைகளில் சீரான வெளிப்பாடு திரும்பத்திரும்ப வருதலையும் அனுமதிக்க

வேண்டும். எடுத்துக்காட்டாக ஆட்டின் மொழியை எடுத்துக் கொண்டால் பின்வரும் கோர்வைகள் அமைந்திருக்கும்.

baa!
baaa!
baaaa!
baaaaa!
baaaaaa!

b என்பதைத் தொடர்ந்து குறைந்தது இரண்டும் 'a' வர இறுதியில் ஆச்சரியக்குறி வரும் கோர்வை இம்மொழியில் அமைந்திருக்கும். திரும்பத்திரும்ப வருவதைக் குறிப்பிட நாம் க்ளீன் ஸ்டார் (Kleene Star*) என்று பொதுவாக அழைக்கப்படும் நட்சத்திரக் குறியீட்டின் (*asterisk*) அடிப்படையில் அமைந்த ஒரு குழும (Set of Operators) இயக்கிகளைப் பயன்படுத்தலாம். க்ளீன் ஸ்டார் "முன்வரும் எழுத்து பூஜிய தடவையோ பலதடவையோ வருதல்" என்பதை உணர்த்தும். எடுத்துக்காட்டாக /a*/ என்பது 'a' பூஜிய தடவையோ பல தடவையோ வரும் கோர்வை" என்று பொருள்படும். இது a என்பதோடும் aaaa என்பதோடும் பொருத்தமுறும். மட்டுமன்றி பூஜிய தடவை a வரும் "Often" என்பதோடும் பொருத்தமுறும். எனவே ஒரு தடவையோ பல தடவையோ a வருவதைக் குறிப்பிடும் சீரான வெளிப்பாடு /aa*/ என்பதாகும். மிகச் சிக்கலான அமைப்பொழுங்கையும் திரும்பக்கூற இயலும். எடுத்துக்காட்டாக /(ab)*/ என்பது "பூஜியம் அல்லது பல தடவை a- க்கள் அல்லது b- க்கள் வரும் என்று பொருள்படும். இது aaaa அல்லது ababab அல்லது bbbb என்ற கோர்வைகளுடன் பொருத்தமுறும்.

இன்டிஜர் (Integer) (ஸ்தான எண்களின் கோர்வை) என்பதன் சீரான வெளிப்பாடு /(0-9)(0-9)*/ என அமையும்.

சீரான வெளிப்பாட்டில் குறைந்தது ஒரு தடவை என்று குறிப்பிட வேண்டி வரும். க்ளினெட் (Klneenet) என்பதைப் பயன்படுத்தினால் முந்திய எழுத்து ஒரு தடவையோ பல தடவையோ" என்று பொருள்படும். எடுத்துக்காட்டாக /(0-9)+/ என்பது "ஸ்தான எண்களின் கோர்வை என்பதைக் குறிப்பிடும் சாதாரண வழியாகும். இவ்வாறு ஆட்டின் மொழியை இரு வகைகளில் குறிப்பிடலாம். /baaa*!/ அல்லது /baa+!/

மிக முக்கியமான பிரத்தேகமான எழுத்து முற்றுப்புள்ளி (முழுப்புள்ளி) ஆகும். இது எந்த ஒரு எழுத்துடனும் பொருத்தமுறும் ஏகோபித்த ஒன்றாகும் (Wild card)

		பொருந்தும்

சீரான வெளிப்பாடு	பொருத்தம்	அமைப்பொழுங்குகளின் எடுத்துக்காட்டு
/beg.n/	Any character between beg after and n	begin, beg'n, began, begun

ஏகோபித்த எழுத்து (.) கிளின் ஸ்டாருடன் சேர்ந்து "எந்த எழுத்துகளின் கோர்வை" என்ற பொருளில் பயன்படுத்தப்படும்.

நங்கூரங்கள் (Anchors)

ஒரு குறிப்பிட்ட இடங்களில் நங்கூரங்கள் சீரான வெளிப்பாட்டை ஒரு கோர்வையின் குறிப்பிட இடங்களில் நிறுவும் தனிப்பட்ட எழுத்துகளாகும். காரட்டும் (Caret=^) டாலர்குறியும் (Dolorsign $) மிகப் பொதுவான நங்கூரங்களாகும். காரட் ^ ஒரு வரியின் ஆரம்பத்துடன் பொருத்தமுறும். எடுத்துக்காட்டாக /^The/ The என்ற சொல்லுடன் பொருத்தமுறும். காரட்டுக்கு இவ்வாறு முன்று பயன்பாடுகள் உண்டு.

1. வரியின் தொடக்கத்துடன் பொருத்தமுறும்.
2. செவ்வக அடைப்புக்குறிக்குள் மறுதலிக்கும்.
3. காரட் என்பதைக் குறிப்பிடும்.

டாலர் குறி ($) ஒரு வரியின் இறுதியுடன் பொருத்தமுறும். எடுத்துக்காட்டாக $ என்ற அமைப்பொழுங்கு வரியின் இறுதியில் வரும் இடைவெளியுடன் பொருத்தமுற பயன்படும். /^The dog\.$/ என்பது The dog என்ற தொடர் உள்ள வரியுடன் மட்டுமே பொருத்தமுறும். (.) என்பது முற்றுப்புள்ளி தான் ஏகோபித்த எழுத்தல்ல என்பதைக் குறிப்பிட இடச் சாய்கோடு (Slash) பயன்படுத்தப்படுகின்றது.

மேலும் இரண்டு நங்கூரங்கள் உண்டு. \b என்பது சொல் எல்லையுடன் பொருத்தமுறும். \B சொல் எல்லை அல்லாமையுடன் பொருத்தமுறும்.

எடுத்துக்காட்டாக/\b The \b/ என்பது The என்ற சொல்லுடன் பொருத்தமுறும்; Other என்ற சொல்லுடன் பொருத்தமுறாது. நுணுக்கமாக பெர்ல் ஒரு சொல்லை இலக்க ஸ்தான எண்கள் (Digits) அல்லது கீழ்க்குறியீடுகள் (Underscores) அல்லது எழுத்துகள் (Letters) என்பதன் வரிசை என்று விளக்குகின்றது. இது பெர்ல் 'சி' என்ற வழியமைப்பு மொழிகளில் (Programing Language) 'Word' என்பது விளக்கப்பட்டுள்ளதை அடிப்படையாகக் கொண்டது. எடுத்துக்காட்டாக /\b99/ என்பது There are 99 bottles என்ற வாக்கியத்திலுள்ள 99 என்ற கோர்வையுடன் தொடர்புறும்; ஆனால் There are 299 bottles என்ற வாக்கியத்திலுள்ள 99 என்ற

கோர்வையுடன் தொடர்புறாது; ஆனால் $ 99 என்பதிலுள்ள 99- உடன் தொடர்புறும். (&- ஒரு ஸ்தான எண்ணோ அடிக் குறியீடோ எழுத்தோ அல்ல)

பிரித்தல் (Disjanction) குழுமுதல் (Grouping) மற்றும் முன்நிலை (Precedence)

பிரித்தலுக்கு பைப் குறியீடு (Pipe Symbol 1) என்ற பிரித்தல் இயக்கி பயன்படுகின்றது. எடுத்துக்காட்டாக /Cat \dog/ என்ற அமைப்பொழுங்கு Cat அல்லது Dog என்ற கோர்வையுடன் பொருத்தமுறும். பிரித்தல் குறியீடு ஒரு குறிப்பிட்ட அமைப்பொழுங்கில் மட்டும் செயலுற பிரை (வில்) அடைப்புக்குறி இயக்கிகளைப் பயன்படுத்த வேண்டிவரும். எடுத்துக்காட்டாக /gupp (Lies)/ என்ற அமைப்பொழுங்கு பிரித்தல். y-க்கும் ies-க்கும் இடையே தான் செயல்படும் என்பதைக் குறிப்பிடும். ஒரு இயக்கி வேறு ஒரு இயக்கிக்கு முன்னர் செயல்படும் என்பது சீரான வெளிப்பாட்டின் இயக்கி முன்னிலைப்படுகை படிநிலை அமைப்பு என்பதால் குறிப்பிடப்படும். பின்வரும் அட்டவணை கூடுதல் முன்னிலைப்படுகையிலிருந்து குறைந்த முன்னிலை படுகை வரை சீரான வெளிப்பாட்டு இயக்கி முன்னிலைப் படுகையின் நிரக்கைத் தரும்.

பிறை அடைப்புக் குறிகள்	( )
எண்ணிகள்	*+? { }
வரிசைகளும் நங்கூரங்களும்	The ^ my end $
பிரிதல்	1

எடுத்துக்காட்டாக எண்ணிகளுக்கு வரிசைகளை விட கூடுதல் முன்னிலைப்படுகை இருப்பதால், /The */ என்பது ‘theeeee’ என்பதுடன் பொருத்தமுறும்; The the என்பதுடன் பொருத்தமுறாது.

எளிதான எடுத்துக்காட்டுகள்.

(i) The என்பதைக் கண்டுபிடிப்பதற்கான சீரான வெளிப்பாடு /the/

(ii) The அல்லது the கண்டுபிடிப்பதற்கான சீரான வெளிப்பாடு /(Tt)he/

(iii) The முன்னும் பின்னும் எழுத்துக்கள் வரவில்லை என்பதைக் குறிப்பிட / / \b [Th] he \b/

(iv) The-இன் முன்னும் பின்னும் எழுத்துக்கள் வரவில்லை என்பதைக் பின்வருமாறு குறிப்பிடலாம்.

/(^a-zA,Z) (tT) he (^a-z A-Z)/

/(^a-zA,Z) (tT) he (^a-z A-Z)/

உயர்நிலை இயக்கிகள் (Advanced Operators)

சீரான வெளிப்பாடு	விரிவு	பொருத்தம்	அமைப்பொழுங்குகளின் எடுத்துக்காட்டு
\d	(0-9)	எந்த ஸ்தான எண்	party of 8
\D	(^0-9)	எந்த ஸ்தான எண்ணும் அல்ல	Blue moon
\w	(a-zAz0-9 g	எந்த எழுத்து எண் அல்லது இடைவெளி அல்ல	Daiyu
\W	(^\W)	எந்த எழுத்து எண் அல்லது இடைவெளி அல்ல	!!!!
\ s	(g /r/t/n/	வெள்ளை இடைவெளி அல்ல	in concord
\ S	(^\s)	வெள்ளை இடைவெளி அல்ல	இடைவெளி தட்டு

சீரான வெளிப்பாடு	பொருத்தம்
*	முன்வரும் எழுத்த பூஜியமோ அதற்கு மோலோ தடவைகள் வருதல்
+	முன்வரும் எழுத்து ஒன்றோ அதற்கு மேலோ தடவைகள் வருதல்
?	முன்வரும் எழுத்த சரியாக பூஜியமோ ஒரு தடவையோ வருதல்
{n}	முன்வரும் எழுத்து n தடவைகள் வருதல்
{n,m}}	முன்வரும் எழுத்து n முதல் m தடவைகள் வருதல்
{n,}	முன்வரும் எழுத்து குறைந்தது n தடவைகள் வருதல்

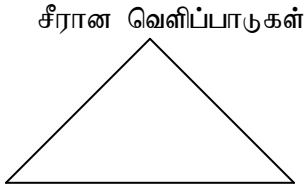
சீரான வெளிப்பாடு	பொருத்தம்	பொருந்தும் அமைப்பொழுங்குகளின் எடுத்துக்காட்டு

*	ஆஸ்டரிக்ஸ் நட்சத்திரக் குறியீடு '*'	"K*A*P*L*A*N"
\.	முழுப்புள்ளி " ."	"Dr. Livingston"
\?	கேள்விக்குறி	"should you light my coedle?"
\n	புது வரி	
\t	ஒரு தட்டு	

முற்றுநிலைத் தானியங்கி (Finite State Automata FSA)

சீரான வெளிப்பாடு பனுவல் தேடலுக்கு வசதியான கருவிமொழி என்பதிலும் மேற்பட்டதாகும். முதலாவதாகச் சீரான வெளிப்பாடு ஒரு முற்றுநிலைத் தானியங்கியை விளக்குவதற்கான ஒரு வழியாகும். இறுதிநிலைத் தானியங்கிகள் கணிக்கும் வேலையின் முக்கிய காரியங்களின் கோட்பாடு அடித்தளமாகும். (நினைவுப் பண்புக்கூறைப் பயன்படுத்தும் சீரான வெளிப்பாடு தவிர) எந்த சீரான வெளிப்பாட்டையும் இறுதிநிலைத் தானியங்கிகளாக நிறைவேற்றலாம். அதே போல் எந்த முற்றுநிலைத் தானியங்கியையும் சீரான வெளிப்பாட்டால் விளக்கலாம்.

இரண்டாவதாக சீரான வெளிப்பாடு என்பது சீரான மொழி (Regular Language) என்று அழைக்கப்படுகின்ற ஒரு வகைக் குறிப்பிட்ட முறையான மொழியை (Formal Language) பண்பாக்கம் செய்யும் ஒரு வழியாகும். சீரான வெளிப்பாடுகளையும் இறுதிநிலை தானியங்கிகளையும் முறையான மொழியை விளக்கப் பயன்படுத்தலாம். இந்த மூன்று கோட்பாட்டுக் கட்டமைப்புகளுக்கு (Three Theoretical Constructions) இடையே உள்ள தொடர்பைப் பின்வரும் படத்தால் காட்டலாம்.



முற்றுநிலைத் தானியங்கி சீரான வெளிப்பாடுகளை நிறைவேற்றப் பயன்படுத்தப்படும் ஒரு கணித உபாயம் (Device) மட்டுமன்றி கணித மொழியியலின் முக்கியக் கருவிகளில் ஒன்றாகும்.

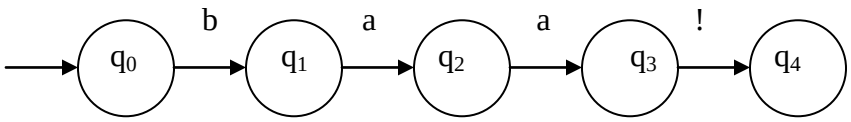
தானியங்கிகளின் வேறுபாடுகளான முற்றுநிலை மாறிகள் (Finite State transducers), ஹிடன் மார்க்கவ் மாதிரிகள் (Hidden Markov Models) மற்றும் N கிராம் இலக்கணங்கள் (N.Gram grammers) என்பன பேச்சைத் தெரிந்து கொள்ளுதல் (Speech recognition) பேச்சு இணைப்பாக்கம் (Speech synthesis) எழுத்துத் திருத்தம் (Spell Checking) செய்திப் பிரித்தெடுப்புப் பயன்பாடுகள் (Information Extraction Applications) என்பனவற்றின் முக்கியமான பகுதிகளாகும்.

இறுதி நிலைத் தானியங்கிகளைக் கொண்டு ஆட்டுமொழியைப் (Sheep talk) புரிந்து கொள்ளுதல்.

ஆட்டு மொழியில் பின்வரும் குழுமத்தில் உள்ள ஏதாவது ஒரு கோர்வை இருப்பதாகக் கருதலாம்.

ba!
baa!
baaa!
baaaa!
baaaaa!

இம்மாதிரியான ஆட்டுமொழியின் சீரான வெளிப்பாடு /baa!+!/ என்று அமையும். பின்வரும் படம் இந்த சீரான வெளிப்பாட்டின் முற்று நிலை தானியங்கி மாதிரியாக அமையும்.



முற்று நிலைத்தானியங்கி ஒரு குழுமம் கோர்வைகளைப் புரிந்துக் கொள்கிறது. இங்கு சீரான வெளிப்பாடு செய்வது போன்று ஆட்டுமொழியைப் பண்பாக்கம் செய்யும் கோர்வைகளைப் புரிந்து கொள்கிறது.

கணுக்களின் முற்றுக் குழுமங்கள் (Finite vertices/nods) மற்றும் இரு கணுக்களை இணைக்கும் விற்கள் மூலமும் நாம் தானியங்கிளையத் திசைநோக்கு வரைபடமாக (Directed graph) உருப்படுத்தம் செய்கிறோம். நாம் கணுக்களை வட்டங்களாகவும் விற்களை அம்புக் குறிகளாலும் உருப்படுத்தம் செயகிறோம். தானியங்கிக்கு ஐந்து நிலைகள் உண்டு. அவை வரைப்படத்தில் கணுக்களாக உருப்படுத்தம் செய்யப்பட்டுடள்ளன. பூஜியநிலை தொடக்க நிலையாகும். இது உள்வரு அம்புக்குறியாக உருப்படுத்தம் செய்யயப்பட்டுள்ளது. நாலாவது நிலை இறுதிநிலை அல்லது கொள்ளும் நிலை (Acceptiong strate) ஆகும். இது இரட்டை

வட்டத்தால் (Double Cricle) உருப்படுத்தம் செய்யப்பட்டுள்ளது. இதில் நான்கு மாற்றங்கள் உண்டு அதை நாம் வரைப்படத்தில் விற்களால் உருப்படுத்தம் செய்துள்ளோம்.

இறுதித் தானியங்கியைக் கோர்வைகளைப் புரிந்து கொள்ள பின்வருமாறு பயன்படுத்தலாம்.

கட்டங்களாகப் பிரிக்கப்பட்டுள்ள ஒரு நீளமான நாடாவில் கட்டத்திற்கு ஒரு குறியீடாக எழுதப்படுவதாகக் கொள்ளலாம்.

தானியங்கி q_0 என்ற தொடக்க நிலையில் தொடங்குகின்றது. பின்வரும் செயல்பாட்டைத் திரும்பத்திரும்பச் செய்கின்றது. உள்ளீட்டின் அடுத்த எழுத்தைப் பரிசோதனை செய்கின்றது. அது வில் குறியின் மேலுள்ள குறியீட்டுடன் பொருத்த முற்றால் தற்போதைய நிலையிலிருந்து வில்லைத் தாண்டி அடுத்த நிலைக்குச் செல்கிறது. மட்டுமல்லாமல் உள்ளீடு இல்லாத ஏற்பு நிலையில் (Acceping state q_4) இருந்தால் பொறி வெற்றிகரமாக ஆட்டுமொழியின் ஒரு நிகழ்வைப் புரிந்துவிட்டது என்று அர்த்தம். உள்ளீடு இல்லாமலோ வில்லின் மேலுள்ள குறியீட்டுடன் பொருத்தமுறாமலோ பொறி இறுதிநிலைக்குச் செல்லவில்லை என்றாலோ இறுதிக்கு முன்பான நிலையில் தடங்கல் ஏற்பட்டது என்றாலோ நாம் பொறி உள்ளீட்டை தள்ளிவிட்டது என்றோ ஏற்றுக் கொள்ளவில்லை என்றோ கூறலாம். நாம் தானியங்கியை நிலைமாற்றமைவு அட்டவணையாலும் உருப்படுத்தம் செய்யலாம். பின்வருவது மேற்கண்ட இறுதிநிலைத் தானியங்கியின் நிலைமாற்ற அட்டவணையாகும்.

	உள்ளீடு
நிலை	ba !
0	1 Ø Ø
1	Ø 2 Ø
2	Ø 3 Ø
3	Ø 3 4
4	Ø Ø Ø

முக்காற் புள்ளியுடன் (:) கூடிய 4- வது நிலை இறுதிநிலையைக் குறிக்கும். Ø தவறான அல்லது விட்டுப்போன மாற்றத்தைக் குறிக்கும்.

ஒரு இறுதிநிலைத் தானியங்கியைப் பின்வரும் 5 அளவுகளால் (Parameters) விளக்கலாம்.

Q : N நிலைகளின் முற்றுக் குழுமம், $q_0, q_1, \ldots\ldots\ldots\ldots q_N$

Σ : குறியீடுகளின் முற்று உள்ளீட்டு அகர எழுத்து

q_0 : தொடக்கநிலை

F : இறுதி நிலைகளின் குழுமம் F⊆∅

δ (q.i) : நிலைகளுக்கு இடையேயுள்ள மாற்றமைவுச் செயல்பாடு அல்லது மாற்றமைவு வார்ப்புரு (Matrix) q∈E மற்றும் உள்ளீட்டுக்குறியீடு i∈Σ என்று தரப்பட δ (q.i) என்பது q_1∈Q என்ற புதிய நிலைக்குத் திரும்பும். இவ்வாறு δ என்பது Q X Σ இலிருந்து Q –உக்கு உள்ள தொடர்பு.

மேற்கண்ட படத்தில் உள்ள ஆட்டு மொழித் தானியங்கிக்கு

Q : { q_0, q_1, q_2, q_3, q_4 }

Σ : {a, b, !}

F : {q_4}

δ (q, i) மேற்கண்ட மாற்றமைவு அட்டவணையால் விளக்கப்பட்டுள்ளது. கீழே தரப்பட்டுள்ள நிலைமாற்றமைவு அட்டவணையைப் பயன்படுத்திய ஒரு கோர்வையைப் புரிந்துகொள்வதற்கான வழிமுறை வரைவு செயல்பாடு உறுதிசெய்யப்பட்ட தெரிகை ஏற்றுக்கொள் அல்லது தள்ளிவிடு வரிசைகிரமம் (index) ←-- நாடாவின் தொடக்கம்
நிகழ்நிலை ←--- பொறியின் ஆரம்பநிலை

கண்ணி (Loop)

உள்ளீட்டின் இறுதி நிலை எட்டிவிட்டால் பின்னர் நிகழ் நிலை ஏற்பு நிலை என்றால் பின்னர் ஏற்றுக் கொள்.

விட்டு விடு

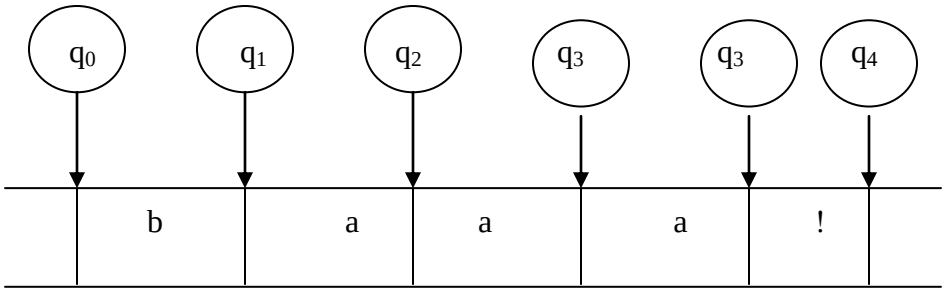
இல்லை என்றால் (Elxif) மாற்றமைவு அட்டவணை நிகழ் நிலை, நாடா (வரிசை) காலி என்றால் பின்னர் விட்டு விடு.
மாறாக நிகழ் நிலை ⟵ மாற்றமைவு அட்டவணை (நிகழ் நிலை, நாடா) வரிசை கிரமம்) வரிசை கிரமம் ⟵ வரிசை கிரமம்

முடிவு :

உறுதி செய்யப்பட்ட தெரிகை ஒரு நாடாவையும் ஒரு தானியங்கியையும் உள்ளீடாக ஏற்கிறது. அது நாடாவில் சுட்டுகிற கோர்வை தானியங்கியால் ஏற்றுக் கொள்ளப்பட்டால் அது ஏற்புக்குத் திரும்புகிறது. இல்லாவிடில் தள்ளிவிடுகிறது.

உறுதி செய்யப்பட்ட தெரிகையானது நாடாவின் தொடக்கத்தில் உள்ள மாறி வரிசைக்கிரமம் மற்றும் பொறியின் தொடக்கநிலையின் தற்போதைய நிலையைத் தூண்டித்தொடங்குகிறது. உறுதிசெய்யப்பட்ட தெரிகை பின்னர் வழிமுறை வரவின் மீதமுள்ளதைச் செயல்படுத்தும் கண்ணிக்குள் செல்கிறது. அது முதலில் அதன் உள்ளீட்டின் இறுதியை அடைந்து விட்டதா என்று பரிசீலிக்கின்றது. அப்படி என்றால் (நிகழ்நிலை ஒரு ஏற்புநிலை என்றால் அது உள்ளீட்டை ஏற்றுக் கொள்ளும் அல்லாவிடில் உள்ளீட்டைத் தள்ளிவிடும்.

நாடாவை உள்ளீடு மீதமிருந்தால் உறுதிசெய்யப்பட்ட தெரிகை எந்த நிலைக்குச் செல்ல வேண்டும் என்பதைத் தீர்மானிக்க மாற்றமைவு அட்டவணையைப் பார்க்கிறது. மாறியான நிகழ்நிலை அட்டவணையிலுள்ள எந்தச் செங்குத்து வரிசையைப் பார்க்க வேண்டும் என்று காட்டும். நாடாவிலுள்ள நிகழ்குறியீடு அட்டவணையிலுள்ள எந்த கிடக்கை வரிசையை பார்க்க வேண்டும் என்பதைக் காட்டும். நேரிடும் கிடைக்கும் மாற்றமைவு அட்டவணைக் கட்டம் மாறியான வரிசைக் கிரமம் நாடாவில் முன்செல்ல உந்தப்படுகின்றது. மாற்றமைவு அட்டவணைக் கட்டம் காலியாக இருந்தால் பொறிக்கு எங்கும் போக இயலாதாகையால் அது உள்ளீட்டைத் தள்ளிவிடுகின்றது. பின்வரும் படம் தரப்பட்ட baaa! என்ற உள்ளீட்டுக் கோர்வையான ஆட்டின் மொழி முற்றுநிலை தானியங்கி மேல் மேற்சொன்ன வழிமுறை வரையின் நிறைவேற்றலைக் காட்டுகின்றது.



நாடாவின் தொடக்கத்தைப் பரிசோதிப்பதற்கு முன் பொறி q_0 என்ற நிலையில் இருக்கிறது. உள்ளீட்டு நாடாவில் b–ஐ கண்டுகொண்டதும் அது மாற்றமைவு அட்டவணையின் உள்ளடக்கத்தில் (q_0,b) சுட்டப்பட்டது போல் q_1 நிலைக்கு மாறுகின்றது. அது பின்னர் a –ஐ கண்டுகொண்டு q_2 நிலைக்கு மாறுகின்றது. மற்றொரு a அதை q_3 நிலைக்குக் கொண்டு செல்கின்றது. முன்றாவது a அதை q_3 நிலையில் விட்டுவிடுகின்றது. அங்கு அது "!" என்பதைக் கண்டுகொண்டு q_4 நிலைக்குச் செல்கின்றது. வேறு உள்ளீடு இல்லாத காரணத்தால் கண்ணியின் தொடக்கத்தில் வரும். (Ex) உள்ளீட்டின் இறுதி என்ற நிபந்தனை முதல் தடவையாக பூர்த்தியாக்கப்பட்டு பொறி q_4 நிலையில் நிற்கும். q_4 என்ற நிலை ஏற்பு நிலை எனவே

பொறியானது baaa! என்ற கோர்வையை ஆட்டுமொழியில் உள்ள ஒரு வாக்கியமாக ஏற்றுக்கொண்டது.

முறையான மொழிகள்

பின்வரும் வரைபடத்தை ஆட்டின் மொழியை ஆக்கும் தானியங்கியாகப் பயன்படுத்தலாம். இதற்கு நாம் தானியங்கி q_0 நிலையில் தொடங்கி விற்களைத் தாண்டி புதிய நிலைகளை அடைந்து அது தாண்டி ஒவ்வொரு வில்லிலும் உள்ள குறியீடுகளை அச்சிடுவதாகக் கொள்ள வேண்டும். தானியங்கி இறுதி நிலையை அடைந்து நிற்கும். நிலை 3- இல் தானியங்கி a! என்பதை அச்சிடுவதையோ நிலை 4- க்கு செல்வதையோ தெரிந்தெடுக்க வேண்டும். அல்லது a- ஐ அச்சிட்டு விட்டு நிலை 3- க்கு திரும்புவதை தெரிந்தெடுக்க வேண்டும்.

முறையான மொழி ஒரு மொழியின் எல்லா அதற்குரிய கோர்வைகளை மட்டும் தெரிந்துகொள்ளும் மற்றும் ஆக்கம் செய்யும் மாதிரிதான் முறையான மொழியின் விளக்கமாக செயல்படும்.

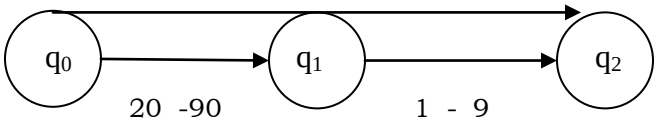
ஒரு முறையான மொழி என்பது கோர்வைகளின் ஒரு குழுமம் ஒவ்வொரு கோர்வைகளும் அகரவரிசை எழுத்து என்று அழைக்கப்படுகின்ற ஒரு முற்று குறியீட்டு குழுமத்திலிருந்து வரும் குறியீடுகளால் ஆனது. ஆட்டின் மொழியின் அகர வரிசை எழுத்து Σ={ a,b,!} என்ற குழுமம் M என்ற மாதிரி எனக் கொண்டால் நாம் L(m) என்பதை "M" என்பதால் பண்பாக்கம் செய்யப்பட்ட முறையான மொழி என்று பொருள்பட பயன்படுத்தலாம். எனவே நாம் முன்னர்பார்த்த ஆட்டின் மொழி தானியங்கியால் விளக்கப்பட்ட முறையான மொழியை பின் வருமாறு முற்றுக் குழுமமாகக் கொள்ளலாம்.

$$L(M) = \{baa!, baaa! ,baaaa!, baaaaaa ! ……..\}$$

ஒரு மூடப்பட்ட வடிவில் முற்றிலாக குழுமத்தை வெளியிடலாம் என்பதுதான் ஒரு மொழியை விளக்க தானியங்கியின் பயன்பாடாகும். முறையான மொழிகள் மக்கள் பேசும் இயற்கை மொழிகள் அல்ல. முறையான மொழி எந்த வகையிலும் உண்மையான மொழியின் ஒப்புமை கொள்ளாமலிருக்கலாம். இருப்பினும் நாம் முறையான மொழியை ஒலியனியல், உருபனியல், தொடரியல் போன்ற இயற்கை மொழியின் பகுதிகளை மாதிரிப்படுத்த பயன்பதுடுத்துகின்றோம். ஆக்கமுறை இலக்கணத்தைக் குறிப்பிடுமாறு பயன்படுத்தப்படுகின்றது. எல்லா இயலக்கூடிய கோர்வைகளை ஆக்கி ஒரு மொழியை விளக்க ஒரு தானியங்கியாகப் பயன்படுவது இந்த கலைச் சொல்லின் மூலமாகும்.

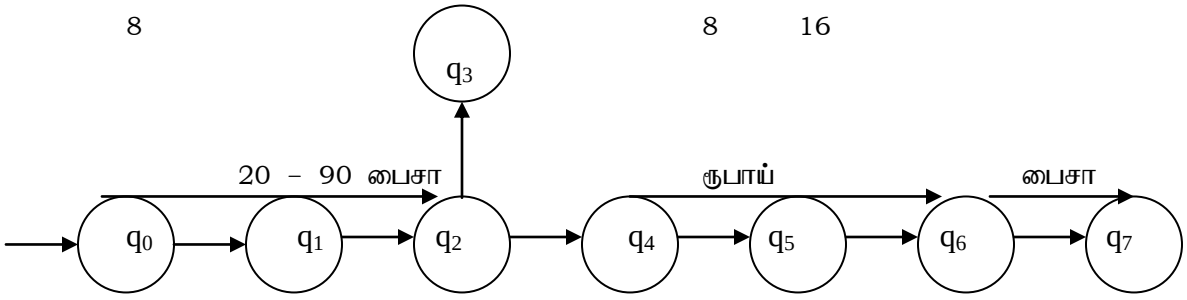
முன்னர் கூறிய எடுத்துக்காட்டுகளில் முறையான அகரவரிசை/நெடுங்கணக்கு எழுத்துக்களால் ஆனது. ஆனால் சொற்களால் ஆன உயர்நிலையிலும் இதை அமல்படுத்தலாம். இவ்வழியில் நாம் சொற்சேர்க்கைப் பற்றிய உண்மகளை மாதிரிப்படுத்திய முற்றுநிலை தானியங்கிகளை எழுதலாம். அம்மாதிரியான முறையான மொழி பத்துபைசா, முன்று ரூபாய், ஒருரூபாய் முப்பது பைசா போன்ற தொடர்களால் ஆன தமிழ்த் துணைக் குழுமங்களை மாதிரிப்படுத்தும்.

1 – 99 வரையிலான தமிழ் எண்களுக்கான முற்று தானியங்கி



1	9	15
2	10	16
3	11	17
4		
5		
6		
7		
8		

1	9	17
2	10	18
3	11	19
4	12	
5	13	
6	14	
7	15	
8	16	



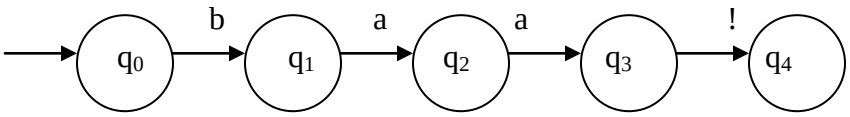
20	1
30	2
40	3
50	4
60	5
70	6
80	7
90	8
	9

20	1
30	2
40	3
50	4
60	5
70	6
80	7
90	8
	9

நாம் ரூபாயின் வேறுபட்ட (நூறு, ஆயிரம், போன்ற) எண்ணிக்கைகளுக்கு இலக்கணத்தில் மேலும் சில சொற்கூறுகளைச் சேர்க்க வேண்டி வரும்.

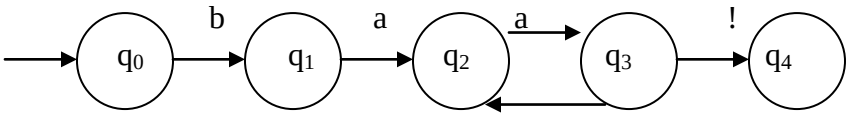
உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கி (Non – determinisitic Finite State Automate = NFSA)

நாம் முன்பு பார்த்த உறுதி செய்யப்பட்ட முற்றுநிலைத் தானியங்கிக்கும், உறுதி செய்யப்படாத தானியங்கிக்கும் உள்ள வேறுபாடு என்னவென்றால் முதல் வகைத் தானியங்கியில் கண்ணி (Loop) நிலை 2-இலும் தற்போதையத் தானியங்கியில் கண்ணிநிலை 3- இலும் உள்ளது.



நாம் 2 –ஆம் நிலையை அடைந்து ஒரு 'a' ஐப் பார்த்தும் நிலை 2 –இல் இருக்க வேண்டுமா அல்லது 3- ஆம் நிலைக்கு செல்ல வேண்டுமா என்று தெரியவில்லை. இம்மாதிரியான நிச்சயிக்கப்பட வேண்டிய முறைகள் உள்ள தானியங்கிகளை உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கி (Non – determinisitic Finite State Automate = NFSA) என்று அழைக்கப்படும்.

குறியீடு இல்லாத விற்களால் ஆன Σ - மாற்றமைவு (Σ - Transition) என்று அழைக்கப்படுகின்ற வேறு ஒரு பொது வகை உறுதிசெய்யப்படாத தன்மையும் இருக்கிறது. பின்வரும் படம் இதை



இந்த தானியங்கி முன்பு பார்த்த அதே மொழியை Σ - மாற்றமைவு மூலம் இறுதியானதாகவே முதலாவதாகவோ விளக்குகிறது. இந்தப் புதியவில்லை பின்வருமாறு பொருள் கொள்ளலாம். நாம் நிலை 3 – இல் இருந்தால் நாம் உள்ளீட்டைப் பார்க்காமால் அல்லது உள்ளீட்டுச் சுட்டை முன்னேற்றி நிலை 2 – க்கு செல்ல அனுமதி உண்டு.

உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கியைக் கோர்வையை ஏற்கப் பயன்படுத்தல் (Using an NFSA to Accept Strings)

நாம் உறுதிசெய்யப்படாத முற்றுநிலைத் தானியங்கியைப் பயன்படுத்தி ஒரு கோர்வை ஆட்டின் மொழியா அல்லவா என்பதை கண்டுகொள்ள, நாம் தவறான வில்லைப் பயன்படுத்தி பின்னர் ஏற்கும் போது விட்டுவிட வேண்டும். ஒரு இடத்தில் ஒன்றிற்கும் மேல் தெரிவு செய்ய சந்தர்ப்பம் இருப்பதால், நாம் தவறானதை தேர்ந்தெடுக்க இயலும். இந்த உறுதி செய்யப்படாத முற்றுநிலைத் தானியங்கி மாதிரிகளின் சிக்கல்களை நாம் கணினி மாதிரி உருவாக்குகையில்

எல்லாம் திரும்பத் திரும்ப நேரிடக்கூடும். இந்தச் சிக்கலுக்கு முன்று நிலைபெறு தீர்வுகள் இருக்கின்றன.

1. திரும்புகை (Backup)

எப்பொழுதெல்லாம் தெரிவு இடத்திற்கு (Cheice point) வருகின்றோமோ நம் உள்ளீட்டில் எங்கிருக்கின்றோம் எந்த நிலையில் தானியங்கி நிற்கிறது என்பதைக் குறிக்க ஒரு அடையாளம் இட இயலும். பின்னர் நாம் தெரிவு செய்தது தவறான வழி என்று புரிந்து கொண்டால் நாம் திரும்பி வந்து வேறு ஒரு பாதையை எடுக்க இயலும்.

2. முன்னோக்கிப் பார்த்தல் (Look ahead)

நாம் உள்ளீட்டில் முன்னோக்கிப் பார்த்து எந்தப் பாதையை எடுப்பது என்பதைத் தீர்மானிக்க உதவியை நாடலாம்.

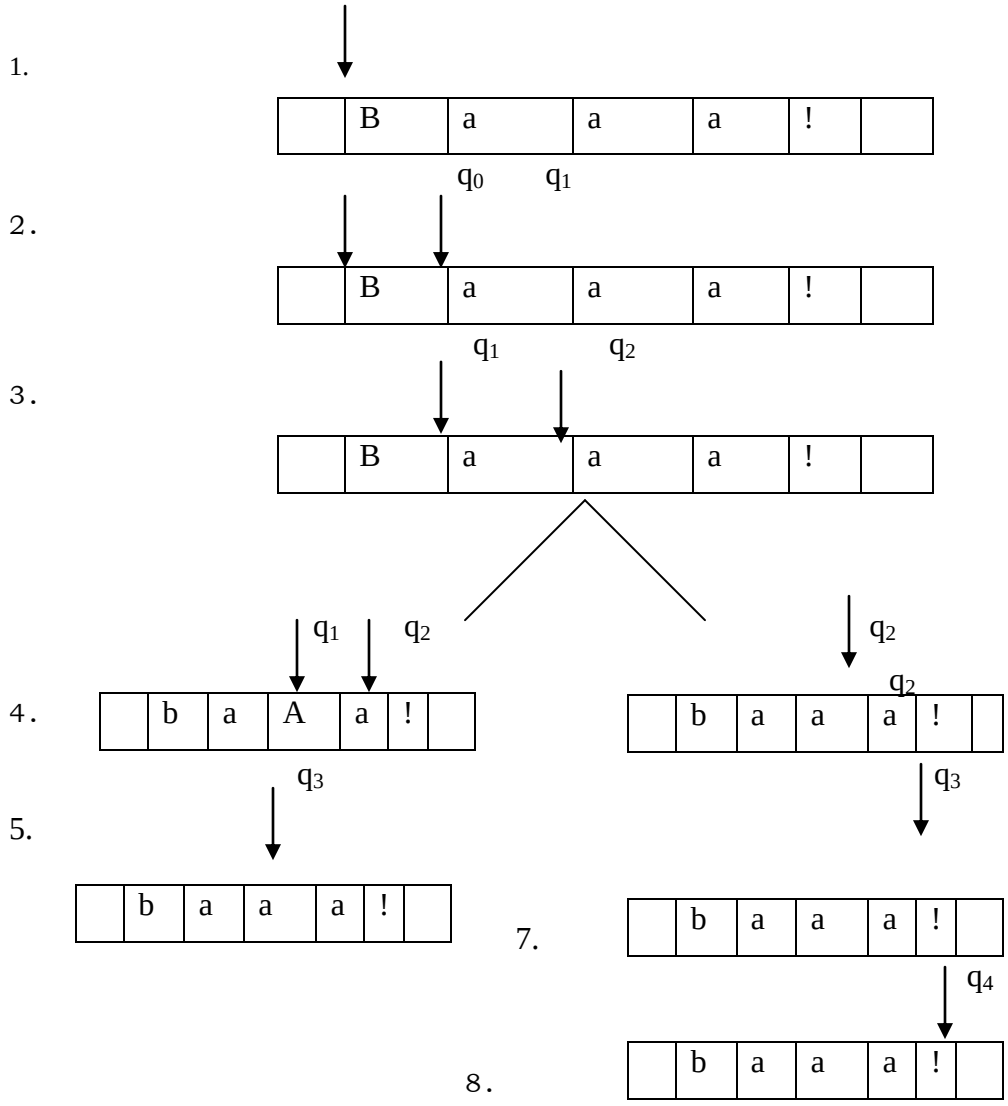
3. இணையம் (Parallelism)

எப்பொழுதெல்லாம் நாம் தெரிவு இடத்திற்கு (Choice point) வருகின்றோமோ நாம் இணையாக இருக்கிற ஒவ்வொரு மாற்றுப் பாதையும் பார்க்கலாம்.

கீழ்வரும் ஆட்டின் மொழிக்கான உறுதிசெய்யப்படாத மாற்றமைவு அட்டவணை

	உள்ளீடு
நிலை	b a ! Ɛ
0	1 Ø Ø Ø
1	Ø 2 Ø Ø
2	Ø 2,3Ø Ø
3	ØØ 4 Ø
4	ØØ Ø Ø

பின்வருவது ஆட்டின் மொழியின் உள்ளீட்டைக் கையாளுவதை காட்டும் உறுதி செய்யப்படாத தெரிகையின் எடுத்துக்காட்டாகும்.



தேடலாகத் தெரிந்துகொள்கை (Rreconition as search)

(பார்க்க சுருக்கம்)

1. நிலை – இடத் தேடல் வழிமுறை வரைவுகள் (State space search algorithm): தீர்வுகளுக்கான தேடலில் சீராக இயங்குகிற உறுதி செய்யப்படாத தெரிந்து கொள்கை (ND RECOGNIZE) போன்றவை நிலை – இடத்தேடல் வழிமுறை வரைவுகள் எனப்படும்.
2. தேடல் நடவடிக்கை Search strateg

   தந்திரம்
3. ஆழம் முதல் என்ற தேடல் (Depth – first strategy or Last in first out LIFO) நிகழ்ச்சி நிரல் ஒரு குவியல் (Agenda as –stack)

4. வீதி முதல் என்ற தேடல் (Breadth – First Strategy of First in first out FIFO) நிகழ்ச்சி நிரல் ஒரு வரிசை (Agenda – as – queue)

சுருக்கம்

- மொழியாய்வின் மிக முக்கியமான அடிப்படைக் கருத்து முற்றுத் தானியங்கியாகும்.
- இரு சீரான வெளிப்பாடு என்ற தானியங்கியை அடிப்படையாகக் கொண்ட செயற்கருவி (Practical trol)
- அமைப்பொழுங்கு பொருத்தத்திற்கு (Pattern Matching) சக்தி வாய்ந்த கருவி சீரான வெளிப்பாடாகும்.
- குறியீடுகளை இணைத்தல் (Concatenation of Symbols)

  குறியீடுகளை பிரித்தல் (Disjunction of Symbols ([ ]), /, and .)

  எண்ணிகள் (Counters) (*,+, { n- m }, நிறுவிகள்/நங்கூரங்கள் (Anchors) ^,$)

  மற்றும் முன்வருகை இயக்கிகள் (Precedence Operators) ((,)) என் அடிப்படைச் செயல்பாடுகள்/இயக்கங்கள் சீரான வெளிப்பாட்டில் அடங்கும்.

- எந்த சீரான வெளிப்பாட்டையும் ஒரு முற்றுநிலைத் தானியங்கியாக உருப்படுத்தம் செய்யலாம்.
- சீரான வெளிப்பாட்டின் பாகமாகக் கருதப்படுகின்ற (1, உடன்கூடிய ( ) நினைவு உயர்ந்த (முன்னேற்றமான) இயக்கமாகும். ஆனால் அதை முற்றுநிலைத் தானியங்கியாக உருப்படுத்தம் செய்ய இயலாது.
- ஒரு தானியங்கி உட்படையாக (Implicity) ஒரு முறையான மொழியை (Formal Language) தானியங்கி ஏற்றுக் கொள்கின்ற கோர்வைகளின் குழுமம் என்று விளக்குகின்றது.
- ஒரு தானியங்கி எழுத்துக்கள், சொற்கள் படச்சாயல்கள் (Graphicimages) என்பன உள்ளடக்கிய எந்தக் குறியீடுகளின் குழுமங்களையும் சொற்றொகுதியாகப் (Vocabulary) பயன்படுத்தலாம்.
- உறுதி செய்யப்பட்ட தானியங்கி (DFSA) அது இருக்கிற நிலையால் முழுவதுமாகத் தீர்மானிக்கப்படுகின்றது.
- உறுதி செய்யப்படாத தானியங்கி (NFSA) நிகழ்நிலையும் (Current state) அடுத்த உள்ளீடு (Next input) தரப்படும் போது அது எடுக்க வேண்டிய பல் வழிகளில் (Multiple Paths) ஒன்றைத் தேர்ந்தெடுக்க வேண்டும்.

- உறுதி செய்யப்படாத தானியங்கியையும் உறுதி செய்யப்பட்ட தானியங்கியாக மாற்ற இயலும்.
- நிகழ்ச்சி நிரலில் (agena) எந்த உறுதி செய்யப்படாத தானியங்கியைத் தெரிந்தெடுக்க வேண்டும் என்ற நிரக்கைப் பரிசோதிப்பது (explore) அதன் தேடல் நடவடிக்கையை/ தந்திரத்தை (Search Stradege) விளக்குகின்றது. ஆழம் முதல் என்ற தேடல் (Depth – firsrt Search LIFO) நிகழ்ச்சி நிரல் ஒரு குவியல் (agenala –as-stack) என்பதுடன் பொருத்தமுடையது (Corrcsponds) வீதி முதல் என்ற தேடல் (Breath first search அல்லது FIFO) என்ற தேடல் நிகழச்சி நிரல் ஒரு வரிசை (agenda – as – gueue) என்பதுடன் பொருத்தமுடையது.
- எந்த சீரான வெளிப்பாட்டையும் தானாகவே உறுதி செய்யப்படாத தானியங்கியாகவும் உறுதி செய்யப்பட்ட தானியங்கியாகவும் தொகுக்க இயலும்.

LIFO: Last In First Out

FIFO: First In First Out

உருபனியலும் முற்றுநிலை மாற்றிகளும் (Morphology and Finite State Transducers)

ஒரு சொல்லின் கூறுகளை ஆயும் செயல்பாடுதான் உருபனியல் பகுத்துக் குறித்தல் (Morpholocal Parsing) எனப்படும்.

எடுத்துக்காட்டு

மரங்கள் ⟶ (மரம்) N + (கள்) PL

மரங்கள் என்பது மரம் மற்றும் கள் என்ற இரு உருபன்களாகப் பிரியும் என்பதைத் தெரிந்து கொள்ளும் செயல்பாடுதான் உருபனியல் பகுத்துக் குறித்தல் எனப்படும். ஒன்றை உள்ளீடாக எடுத்துக் கொண்டு ஒரு வித அமைப்பை அதற்குத் தருவதைப் பொதுவாகப் பகுத்துக் குறித்தல் (Parsing) எனலாம்.

ஆட்சி செய்தி (Information Retrival) மீட்பு பரப்பில் (Domain) மரங்கள் என்பதிலிருந்து மரம் என்பதைப் பொருத்திப் பார்கும் (Morping) சிக்கலை பகுதியாக்கம் (Stemming) என்பார்கள். உருபனியல் பகுத்துக் குறித்தலோ பகுதியாக்கம் செய்தாலோ பன்மை ஒட்டுகளுக்கு மட்டுமன்றி வேறு பல ஒட்டுகளுக்கும் பயன்படும். எடுத்துக்காட்டாக வந்தான் என்பதை (வ) N+ (ந்த்) PAST + (ஆன்) TMS என்று பகுத்துக்குறிக்கலாம் வ என்பது வினைப் பகுதி 'ந்த்' என்பது இநற்ததால ஒட்டு ஆன் என்று பகுத்துக் குறிக்கப்படும்.

நாம் தமிழ்ப் பெயர்களின் எல்லாம் பண்டை வடிவுகளையும் அகராதியல் பட்டியலிட்டுத் தருவதில்லை ஏனென்றால் கள் என்ற பன்மைப் பின் ஒட்டு விளைவாக்கமுள்ளது (prodnctive). இது எல்லாப் பொயர்களுடனும் சேர்ந்து பன்மைப் பெயர் வடிவங்களை உருவாக்கும் தன்மை உடையது. எனவே வினைகள் மற்றும் பெயர்களின் எல்லாத் திரிபு வடிவுகளையும் பட்டியலிடுவது திறமையற்ற செயல் (அதாவது சிக்கலானதல்ல) மேலும் விளைவாக்கம் உள்ள பின்னொட்டுகள் புதிய சொற்களுடனும் பயன்படுத்தப்படும். (எடுத்துக்காட்டாக கணினி – கணினிகள்) புதிய சொற்கள் நாள்தோறும் உருவாக்கப்படுகின்றன. இவற்றிற்கெல்லாம் பன்மை வடிவுகளைப் பட்டியலிட்டுக் கொண்டே செல்ல இயலாது. மேலும் துருக்கிய மொழி (Turkish Language) போன்ற சிக்கலான உருபனியல் அமைப்புள்ள மொழிகளின் சொற்களின் எல்லா வடிவுகளையும் சேகரித்து வைக்க இயலாது.

உருபனியல்

குறைந்த பொருள்தருகின்ற அலகுகளான உருபன்களால் சொற்கள் எவ்வாறு உருவாக்கப்படுகிறது என்ற பாடம் உருபனியலாகும். உருபன் பொருண்மையைக் கொண்டிருக்கின்ற குறைந்த மொழி அலகாகும். எடுத்துக்காட்டாக 'கால்' என்ற சொல்லில் ஒரு உருபன் உள்ளது. கால்கள் என்ற சொல்லில் கால், கள் என்ற இரு உருபன்கள் உள்ளன.

உருபன்களைப் பகுதிகள் (Stems) ஒட்டுகள் (Affxes) என்ற இரண்டு பெரிய உருபுகளாகப் பிரிக்கலாம். இவற்றிற்கு இடையே உள்ள வேறுபாடு மொழிக்கு மொழி வேறுபாடும் உள்உணர்வின் அடிப்படையில் பகுதி என்பது முக்கியப் பொருள் தரும் சொல்லின் முக்கியமான உருபன் என்றும் ஒட்டுகள் என்பது வேறுபட்ட வகையிலான கூடுதல் பொருள் ஊட்டும் உருபன்கள் என்றும் கூறலாம்.

ஒட்டுகள் முன்னொட்டுகள் (Prefixes) பின்னொட்டுகள் (Suffixes) இடையொட்டுகள் (Infixes) மற்றும் ஓர ஒட்டுகள் (Circumfixes) என்று பகுக்கலாம். முன்னொட்டுகள் பகுதிக்கு முன்வரும் பின்னொட்டுகள் பகுதிக்கு பின்னுமாக வரும். இடையொட்டுகள் பகுதிக்குள் சொருகப்படும். எடுத்துக்காட்டாக தமிழில் மான்கள் என்ற சொல்லில் வரும் பன்மை ஒட்டான (உருபன்) கள் பின்னொட்டாகும். அநீதி முன்னொட்டாகும். அநியாயம் என்ற சொற்களில் வரும் அ என்பது முன்னொட்டாகும். தமிழிலும் ஆங்கிலத்திலும் ஓர ஒட்டுகள் இல்லை. ஜெர்மன் மொழியில் ஓர ஒட்டுகள் உண்டு.

எடுத்துக்காட்டு

Film - en ' to film' – ge – film - t ' filemed'

Frag - en ‘to ask’ _ ge _ frag -t ‘asked’

Lob - en ‘to praise’ _ ge _ Lob - t ‘Praisd’

Zeig - en to ‘Show’ _ ge _ Zeig - en ‘Shown’

ge யும் to யும் இணைந்து Past Participle (இநற்தகால எச்சம்) வடிவை உருவாக்கும். பிலிப்பைன் மொழியான டாகலாக் (Tagalog) வியட்நாம் மொழியான சரவ் (Charau) போன்ற மொழிகளில் இடை ஒட்டுகள் பயன்படுத்தப்படுகின்றன.

சரவ் மொழியிலிருந்து (எ.கா)

Sulat ‘Write’ – S – um – Ulat ‘Wrote’

S – in – Ulat ‘Written’

டாகலாக் மொழியிலிருந்து (எ.கா)

Hingi ‘borrow’ – h – um – ingi

ஒரு சொல் பல உருபன்களின் பிணைக்கப்பட்டுக் உருவாக்கப்படுவதால் முன் ஒட்டுகளும், பின்னொட்டுகளும் பிணைப்பு உருபனியல் (Concat enative Morphology) எனப்படும். பல வழிகளில் உருபன்கள் மிகச் சிக்கலான வழியின் இணையும் பிணைப்பில்லாத உருபனியல் (Non – Concutenative Morphology) பரவலாக இருக்கும். மேற்சொன்ன டாகலாக் மொழி பிணைப்பு உருபனியலுக்கு எடுத்துக்காட்டாகும். வேறு ஒருவகையான பிணைப்பில்லாத உருபனியல் வார்ப்புச் சட்ட உருபனியல் (Templatic Morphology) அல்லது வேர் அமைப்பொழுங்கு உருபனியல் (root – Pattern Morphology) எனப்படும். அரேபிய மொழி (Arabic Language) ஹீப்ரு மொழி பிற செமிட்டிக் மொழிகளில் (Semitic Language) சாதாரணமாக வார்ப்புச் சட்ட உருபனியல் காணப்படுகிறது. ஹீப்ரு மொழியில் வினை இரண்டு உட்கூறுகளால் உருவாக்கப்படுகின்றது. ஒன்று அடிப்படைப் பொருளைக் கொண்டிருக்கும் மூன்று மெய்களால் (CCC) ஆன வேர், (வார்புச் சட்டம் இரண்டு பொதுவாக வினைப்பாட்டுப் (Voice) பொருளை வெளிப்படுத்தும் உயிர்கள் எடுத்துக்காட்டாக மூன்று மெய்யால் ஆன Learn அல்லது ‘Study’ என்று பொருள் தருகின்ற Lnad என்ற வேர் செய்வினையைக் (Active Voice) குறிக்கும் CaCaC என்ற வார்புச் சட்டத்துடன் சேர்ந்து Lamad ‘he studied’ ‘அவன் படித்தான்’ என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தலைக் குறிக்கும். Cicec என்ற வார்புச் சட்டத்துடன் சேர்ந்து Limed ‘he Laught’ அவன் கற்பித்தான்’ என்ற சொல்லை உருவாக்கும் அல்லது வலியுறுத்தல் செயப்பாட்டு வினையைக் குறிக்கும். cacac என்ற வார்புச்

சட்டத்துடன் சேர்ந்து Lumad 'he was Laught' அவன் கற்பிக்கப்பட்டான் என்ற சொல்லை உருவாக்கும்.

ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட ஒட்டுக்கள் வரலாம். Ex. பானைகளை என்ற சொல்லில் பானை என்ற பகுதியும் கள் என்ற பன்மை விகுதியும் ஐ என்ற 2-ஆம் வேற்றுமை உருபும் உள்ளன. Rewrites என்ற சொல்லல் re முன்னொட்டும் write என்ற பகுதியும் s என்ற பின்னொட்டும் உள்ளன. Unbelievably என்ற ஆங்கிலச் சொல்லில் Un என்ற முன்னொட்டும் belive என்ற பகுதியும் able, ly, என்ற பின்னொட்டுக்களும் உள்ளன. தமிழ் போன்ற மொழிகளில் மூன்றோ நான்கோ ஒட்டுக்களை கொண்டிருக்கும். ஆனால் துருக்கிய மொழி போன்ற மொழிகளில் சொற்கள் ஒன்பதோ பத்தோ ஒட்டுக்களைக் கொண்டிருக்கும். இம்மாதிரியான மொழிகள் ஒட்டுக்கள் நிறைந்த மொழிகள் (agglutinative language) என்று அழைக்கப்படும்.

உருபன்களிலிருந்து சொற்கள் உருவாக்கப்படுவதை இரண்டு பெரும் வகுப்புகளாகப் பகுக்கலாம். அவைகள் திரிபுகள் (Inflection) ஆக்கங்கள் (Derivation) என்பன. ஒரு சொல் பகுதி ஒரு இலக்கண உருபனுடன் இணைந்து அதே சொல்வகுப்பைச் சேர்ந்த சொல்வடிவை விளைவிப்பது திரிபு எனப்படும். Ex. பன்மை ஒட்டான கள் ஒருமைப் பெயர்ப்பகுதியுடன் இணைந்து பன்மை இலக்கண உருபனுடன் இணைந்து பொதுவாக வேறு ஒரு வகுப்பைச் சார்ந்த பெரும்பாலும் பொருளைச் சிக்கலான சொல்வடிவை உருவாக்குவது ஆக்கம் (Derivation) எனப்படும்.

இயல் 3

சொற்களை ஆய்தல்

தமிழ் உருபனியல் (Tamil Morphology)

தமிழ் உருபனியல் முழுவதுமான/நிறைவான பிணைப்பு உருபனியல் (Concatenative Morphology) எனலாம். இதன்படி உருபுகள் எல்லாம் தொடர்ச்சியாக ஒட்டப்பட்டுள்ளன. தமிழ் பிணைப்பு உருபனியல் பின்னொட்டுகளாக பகுதிக்குப் பின்னர் சேர்க்கப்பட்டுள்ளது. இதைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

பகுதி (+ ஒட்டு) ^{n}

இதில் n பின்னொட்டுகள் ஒன்றோ பல தடவையோ வருவதைக் குறிப்பிடும்.

உருபனியல் மற்றும் தொடரியல் அடிப்படையில் தமிழ்ச் சொற்களை பின்வருமாறு எட்டு சொல் வகுப்புகளாகப் பகுக்கலாம்.

1. பெயர்கள் (Nouns)
2. வினைகள் (Verb)
3. பின்னுருபுகள் (Post Position)
4. பெயரடைகள் (Adjectives)
5. வினையடைகள் (Adverbs)
6. அளவை அடைகள் (Gnontifiers)
7. அடைகொளி அடைகள் (Determiners)
8. இணைப்பிகள் (Conjanction)

பெயரும் திரிபும் (Nouns and Inflection)

ஒரு பெயர் எண்ணுக்காகவும் அதாவது பன்மை எண்ணுக்காகவும் வேற்றுமைக்காகவும் திரிபுறும். திரிபுற்ற ஒரு பெயர் பகுதி, எண், வேற்றுமை என்ற மூன்றும் உருபன்களின் ஒலியனியல் உருப்படுத்தம் எனலாம்.

பெயர் (+எண்) (+ வேற்றுமை)

அக அமைப்பில் இரண்டு இலக்கண உருபன்கள் (அதாவது எண்ணும் வேற்றுமையும்) பெயர் உருபனோடு வரலாம். ஒலியனியல் நிலையில் நான்கு வகையான உருபுகள் (Morphy) பெயர்பகுதியுடன் வரும்.

1. பன்மைப் பின்னொட்டு (Phiral Suffix)
2. திரிபு பின்னொட்டு (Obligue Suffix)
3. ஒலி நிரவல் பின்னொட்டு (Euphonic Surfix)
4. வேற்றுமைப் பின்னொட்டு (Case Surfix)

பெயர் எண்ணிற்காதத் திரிபுறும் போது பன்மைப் பின்னொட்டு பொயர்ப்பகுதியுடன் சேர்க்கப்படும்.

பெயர்ப் பகுதி + பன்மைப் பின்னொட்டு

எடுத்துக்காட்டு

வீடு – கள்

படி – கள்

ஒலி நிரவல் சாரியை (Euphonic Increment) இன், இற்/அன், அற் பெயர்ப்பகுதிக்கும் அல்லது திரிபுப்பகுதிக்கும் வேற்றுமை உருபுகளுக்கும் இடையில் வரும்.

{ பெயர்ப் பகுதி }
அல்லது (+ ஒலி நிரவல் சாரியை) + வேற்றுமை உருபு
{ திரிபுப் பெயர்ப்பகுதி }

எடுத்துக்காட்டு

சாவி – இன் - ஆல்

மர –த்த் - இன் - ஆல்

அத் - அன் - ஆல்

இத் - அன் - ஆல்

அத் - அற் - கு

இத் - அற் - கு

பெயர் எண்ணுக்காகவும் வேற்றுமைக்காகவும் திரிபுற்றால் முதலில் எண் ஒட்டு பகுதியுடன் சேர்க்கப்பட்டு அதைத் தொடர்ந்து விருப்பாக ஒலி நிரவல் வர அதன் பின்னர் வேற்றுமை உருபு சேர்க்கப்படும்.

பெயர்ப் பகுதி + எண் ஒட்டு (+ ஒலி நிரவல் சாரியை) + வேற்றுமை உருபு

திரிபுப் பெயர்பகுதி

திரிபுப் பெயர்பகுதி மூன்று விதமாக உருப்படுத்தம் பெறும்.

1. அம் என முடியும் பெயர்களுடன் ம் கெட அத்து சாரியை சேர்க்கப்படும்.

எ.கா.

	திரிபுப் பெயர்பகுதி + வேற்றுமை உருபு
மர – த்து	மரத்தில்
பழ – த்து	மரத்தை
நில – த்து	மரத்துக்கு
	மரத்தோடு

2. பல. சில, இவை, அவை, எல்லாம் என்ற பெயர்ச்சொற்களில் அற்று சாரியைச் சேர்க்கப்படும் *எல்லாம்* என்பதில் ம் கெட அற்று சாரியைச் சேர்க்கப்படும். இவை, அவை என்பன இ, அ எனச் சுருங்கி அற்றுச் சாரியைச் சேர்க்கப்படும்.

	திரிபுப் பெயர்பகுதி + வேற்றுமை உருபு
பல – வ் - அற்று	பலவற்றில்
சில – வ் - அற்று	பலவற்றை
இவ் - அற்று	பலவற்றுக்கு

அ –வ் - அற்று பலவற்றோடு

எல்லா –வ் - அற்று பலவற்றின்

3. டு. று என முடியும் ஒரு குழுமம் பெயர்களின் இறுதியில் வரும் மெய் இரட்டும்.

திரிபுப் பெயர்பகுதி + வேற்றுமை உருபு

வீடு – வீட்டு	வீட்டில்
நாடு – நாட்டு	வீட்டை
ஆறு – ஆற்று	வீட்டுக்கு
கிணறு – கிணற்று	வீட்டோடு
	வீட்டின்

மாற்றுப்பெயர்களின் திரிபுப்பகுதி

மூவிட மாற்றுப்பெயர்களின் திரிபு வடிவங்கள்

வ.எண்	**மூவிடம்**	**மாற்றுப்பெயர்**	**திரிபுவடிவு**
1	தன்மையிடம்	நான்	என் ~ என
2	முன்னிலையிடம்	நீ	உன் ~ உன
3	ஆண்பால் ஒருமை படர்க்கையிடம்	இவன்/அவன்	இவன்/அவன்
4	பெண்பால் ஒருமை படர்க்கையிடம்	இவள்/ அவள்	இவள்/அவள்
5	தற்சுட்டு படர்க்டைகயிடம்	தான்	தன ~ தன்
6	பன்மை தன்மையிடம்	நாங்கள்/நாம்	எங்கள் நம் ~ நம
7	பன்மை முன்னிலையிடம்	நீங்கள்	உங்கள்
8	ஆண், பெண் பால் பன்மை படர்க்கையிடம்	இவர்கள்/அவர்கள்	இவர்கள்/அவர்கள்
9	ஆ.ினை பன்மை படர்க்கையிடம்	இவை/அவை இவைகள்/ அவைகள்	இவற்றுள்/அவற்றுள் இவைகள்/அவைகள்
10	பன்மை தற்சுட்டு	தாங்கள்	தங்கள்

திரிபுப்பெயர்பகுதி + வேற்றுமை உருபு

என்னை	என்னோடு
எனக்கு	என்னுடன்
என்னில்	என்னிடம்
என்னால்	எனது

எண் திரிபு

பன்மை எண் உருபனியல் அடிப்படையில் பன்மை ஒட்டான கள் - க்கள் என்பதால் அடையாளப்படுத்தப்படுகிறது. கள், க்கள் என்பன பின்வருமாறு வருகைமுறை செய்யும். க்கள் நெடிலீற்றுப் பெயர்கள் முன்னும் உ வில் முடியும் இரண்டு குற்றசைகளைக் கொண்ட பெயர்கள் முன்னும் வரும் பிற இடங்களில் கள் வரும்.

விகுதி விலக்குகள்

பூ - பூக்கள்

ஆண்பால், பெண்பால் , பலர்பால் ஒட்டுகளுடன் கூடிய பகுதிப் பெயர்கள்

பசு _ பசுக்கள்

தெரு _ தெருக்கள்

கல் - கற்கள்

முள் - முற்கள்

மரம் - மரங்கள்

வீடு – வீடுகள்

காசு – காசுகள்

வண்டு – வண்டுகள்

மாணவன் - கள்

மாணவர் - கள்

மாணவி – கள்

நண்பன் - கள்

நண்பர் - கள்

ஒலிநிரவல் சாரியை : இன், அன்

வவர, அற்று என்பன கட்டாயமான ஒலிநிரவல் சாரியைகள் இன் ~ இற், அன் ~ அற் என்பன விரும்ப ஒலிநிரவல் சாரியைகள் அன் என்பது அது என்பதுடன் இணைந்து வரும். இன் பிறவற்றில் வரும். நான்காம் வேற்றுமை உருபு கு முன் ன் என்பது ற் ஆக மாறும்.

வேற்றுமை உருபுகள்

தற்காலத் தமிழில் ஆறு வேற்றுமை பின் ஒட்டுகள் உள்ளன.

வ.எண்	**வேற்றுமை**	**ஒட்டு**
1	செயபடுபொருள் வேற்றுமை	ஐ

2	கொடை வேற்றுமை	உக்கு ~ க்கு ~ கு
3	கருவி வேற்றுமை	ஆல்
4	உடன்நிகழ்ச்சி வேற்றுமை	ஓடு
5	இட வேற்றுமை	இல்

ஐந்து கட்டுண்ட பின்உருபுகள் (Bound Postpossition) உள்ளன.

வ. எண்	**வேற்றுமை**	**பின்னுருபுகள்**
1	இடவேற்றுமை	இடம்
2	உடன்நிகழ்ச்சி வேற்றுமை	உடன்
3	நீங்கல் வேற்றுமை	இருந்து
4	பயனுரு வேற்றுமை	ஆக

வேற்றுமை உருபுகளும் சந்தியும்

1. ஆ.ஒ, ஊ என முடியும் பெயர்களையும் (C) vcv என்ற அசை அமைப்பைக் கொண்டு உ – வில் முடியும் பெயர்களையும் தொடர்ந்து வேற்றுமை உருபுகள் வரும் போது வ் என்ற இடையின ஒலியன் இடையில் வரும்.

ஆ உயிர்தொடக்க ஆ உயிர்தொடக்க
ஒ + வேற்றுமை ஒட்டுகள் ⟶ ஒ -வ்- வேற்றுமைஒட்டுகள்
ஊ ஊ

எ.கா

I. கடா + ஐ – கடாவை, கடா + ஆல் - கடாவால், கடா + ஓடு – கடாவோடு, கடா + உடன் - கடாவுடன், கடா + உக்கு – கடாவுக்கு, கடா + இன் - கடாவின், கடா + இல் - கடாவில் CVC - உ

II. பசு + ஐ – பசுவை, பசு + ஆல் - பசுவால், பசு + ஓடு – பசுவோடு, பசு + உடன் - பசுவுடன், பசு + உக்கு – பசுவுக்கு, பசு + இன் - பசுவின், பசு + இல் - பசுவில்

2. இ,ஈ, ஐ என்று முடியும் பெயர்கள் உக்கு தவிர்த்த பிற வேற்றமை ஒட்டுகள் ஏற்கும் போது ய் என்று இடையின ஒலி இடையில் வரும்.

எ.கா.

எலி + ஐ - எலியை, எலி + ஆல் - எலியால், எலி + ஓடு – எலியோடு, எலி + உடன் - எலியுடன், எலி + இன் - எலியின், எலி + இல் - எலியில்

3. இ, ஈ, ஐ, ய் என்று முடியும் பெயர்கள் முன்வரும் உக்கு வின் உ கெடும்.

எ.கா.

இ		இ
ஈ	+ உக்கு -----→	ஈ - க்கு
ஐ		ஐ
ய்		ய்

எ.கா

எலி + உக்கு > எலிக்கு, ஈ + உக்கு > ஈக்கு, யானை + உக்கு > யானைக்கு, நெய் + உக்கு > நெய்க்கு

4. உ – வில் முடியும் பெயர்களை அவற்றின் சந்திச்செயல்பாட அடிப்படையில் நான்கு வகைகளாகப் பிரிக்கலாம்.

1. (C) VV – ட் - உ என்ற அசை அமைப்பைக் கொண்ட சொற்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமை உருபுகள் வரும்போது உ கெட, ட் இரட்டிக்கும்.

(C) VV, ட் - உ + [உயிர்த்தொக்க வேற்றுமை ஒட்டுகள்] ---→ (C) VV, ட்ட் -→ [உயிர்த்தொடக்க வேற்றுமை ஒட்டுகள்]

எ.கா.

காடு + ஐ – காட்டை, காடு + ஆல் - காட்டால், காடு + ஓடு – காட்டோடு, காடு + உடன் - காட்டுடன், காடு + உக்கு – காட்டுக்கு, காடு + இன் - காட்டின், காடு + இல் - காட்டில்.

2. று – வில் முடியும் பெயர்களைத் தொடர்ந்து உயிர்த் தொடக்க வேற்றுமை உருபுகள் வரும் போது உ கெட ற் இரட்டும்.

று + உயிர்த்தொடக்க வேற்றுமை ஒட்டுகள் ---→ ற்ற – உயிர்த்தொடக்க வேற்றுமை ஒட்டுகள்

(C) VCC – உ (அ) (C) VVC – உ என்ற அசை அமைப்புடைய பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமை உருபுகள் வருகையில் உ கெடும்.

- (C) VCC – உ + [உயிர்த்தொக்க வேற்றுமை ஒட்டுகள்] -→ CVCC - [உயிர்த்தொக்க வேற்றுமை ஒட்டுகள்]

I. வண்டு + ஐ – வண்டை, வண்டு + ஆல் - வண்டால், வண்டு + ஓடு – வண்டோடு, வண்டு + உடன் - வண்டுடன், வண்டு + உக்கு – வண்டுக்கு, வண்டு + இன் - வண்டின், வண்டு + இல் - வண்டில்

II. காது + ஐ – காதை, காது + ஓடு – காதோடு, காது + இல் - காதில், காது + ஆல் - காதால், காது + உடன் - காதுடன், காது + உக்கு – காதுக்கு, காது + இன் - காதின்.

5. (C) VC என்ற அமைப்பையுடைய ண், ய், ல், ன் என முடியும் பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமை பின் ஒட்டுகள் வர முறையே ண், ய், ல், ன், ள் என்பன இரட்டும். (விதி விலக்கு – ய் + உக்கு – ய்க்கு (எ.கா. நெய் + உக்கு – நெய்க்கு

- ண்		ண்ண்
- ய்		ய்ய்
- ல்	+ உயிர்த்தொடக்க வேற்றுமை –	ல்ல் - உயிர்த்தொடக்க வேற்றுமை
- ள்	பின் ஒட்டுகள்	ள்ள் பின் ஒட்டுகள்
- ன்		ன்ன்

I. கண் + ஐ – கண்னை, கண் + ஆல் - கண்ணால், கண் + உக்கு – கண்ணுக்கு, கண் + ஓடு – கண்ணோடு, கண் + உடைய – கண்ணுடைய, கண் + இன் - கண்ணின், கண் + அது - கண்ணது, கண் + இல் - கண்ணில்.

II. நெய்+ஐ – நெய்யை, நெய் + உடைய – நெய்யுடைய, நெய் + ஓடு – நெய்யோடு, நெய் + ஆல் - நெய்யால், நெய் + இன் - நெய்யின், நெய் + உடன் - நெய்யுடன்.

III. கல் + ஐ – கல்லை, கல் + இல் - கல்லில், கல் + கல்லால், கல் + ஓடு – கல்லோடு, கல் + உக்கு – கல்லுக்கு, கல் + அது – கல்லது, கல் + இன் - கல்லின், கல் + உடைய – கல்லுடைய.

IV. முள் + ஐ – முள்ளை, முள் + ஆல் - முள்ளால், முள் + உடன் - முள்ளுடன், முள் + ஓடு – முள்ளோடு, முள் + உடைய – முள்ளுடைய, முள் + இல் - முள்ளில், முள் + இன் - முள்ளின், முள் + அது – முள்ளது.

V. பொன் + ஐ – பொன்னை, பொன் + இல் - பொன்னில், பொன் + உக்கு – பொன்னுக்கு, பொன் + ஆல், பொன்னால், பொன் + ஓடு – பொன்னோடு, பொன் + இன் - பொன்னின், பொன் + உடன் - பொன்னுடன்.

6. அம் ஈற்றுப்பெயர்களைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டுகள் வரும் போதும் கெட த்து என்ற சாரியை இடையில் வரும்.

அம் + [உயிர்த்தொடக்க வேற்றுமை பின்னொட்டு] --→ அ – த்து – [உயிர்த்தொடக்க வேற்றுமை பின்னொட்டு]

எ.கா. மரம் + ஐ – மரத்தை

அது, ஓடு, ஆல், உடன், உடைய, அது, இல்

7. ஆ – ஈற்று பெயர்களான பல, சில என்பனவற்றைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமை பின்னொட்டுகள் வரும்போது அற்று என்ற சாரியை இடையில் வரும்.

அ + [உயிர்த்தொடக்க வேற்றுமை பின்னொட்டு] --→ அ – ற்ற – [உயிர்த்தொடக்க வேற்றுமை பின்னொட்டு]

பல + ஐ – பலவற்றை, பல + இல் - பலவற்றில், உடைய, உடன்

பல + உக்கு – பலவற்றுக்கு, பல + ஆல் - பலவற்றால், அது, ஓடு

8. எல்லாம் என்ற பெயரைத் தொடர்ந்து உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டு வரும்போது ம் கெட அற்று என்ற சாரியை இடையில் வரும்.

எல்லாம் + [உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டு] → எல்லா –வ் அற்று – [உயிர்த்தொடக்க வேற்றுமைப் பின்னொட்டு]

ய் ஈற்றுப் பெயர்களைத் தொடர்ந்து உக்கு என்ற வேற்றுமைப் பின்னொட்டு வரும்போது உக்கு என்பது உ கெட க்கு ஆக மாறும்.

ய் + உக்கு – ய் - க்கு

நாய் + உக்கு – நாய்க்கு, நெய் + உக்கு – நெய்க்கு

10. 5 – ஆம், 6 ஆம் விதிகளுக்கு உட்படாத ண், ய், ல், ள், ன் மற்றும் ர், ழ் என முடியும் பெயர்களைத் தொடர்ந்து உயிர்தொடக்க வேற்றுமைப் பின்னொட்டுகள் வரும்போது அவை எந்த வித மாற்றமுமின்றி இணையும் (விதி விலக்கு : ய் + உக்கு > ய்க்கு

I. தூண் + ஐ, தூணை, தூண் + அது – துணது, தூண் + உடன் - தூணுடன், தூண் + ஆல் - தூணால், தூண் + இன் - தூணின், தூண் + உக்கு – தூணுக்கு, தூண் + ஓடு – தூனோடு.

II. நாய் + ஐ – நாயை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

III. கால் + ஐ – காலை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

IV. வாள் + ஐ – வாளை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

V. மான் + ஐ – மானை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

VI. சுவர் + ஐ – சுவரை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

VII. இதழ் + ஐ – இதழை, உடன், ஓடு, ஆல், அது, இல், இன், உக்கு

11. இன் என்ற சாரியை ஒருமைப் பன்மைப் பெயரடிகளை தொடர்ந்து வேற்றுமை உருபுகளுக்கு முன்னர் விருப்பாக வரும்.

எ.கா. பசு + ஐ – பசுவை

ஆல், இன், இல், உடன், அது, ஓடு, உக்கு

I. த்/கிற்/வ் பிரிவு

இதில் அடங்கும் வினைகளை ஐந்து உட்பிரிவுகளாகப் பாகுபாடு செய்யலாம்.

வ.எண்	**வினைபகுதி**	**இறந்தகால வி.மு. ஆண்பால்**	**நிகழ்கால வி.மு. ஆண்பால்**	**எதிர்கால வி. மு. ஆண்பால்**	**எதிர்கால வி.மு. ஒன்றன்பால்**
1	செய்	செய்தான்	செய்கிறான்	செய்வான்	செய்யும்
2	அழு	அழுதான்	அழுகிறான்	அழுவான்	அழும்
3	சாகு	செத்தான்	சாகிறான்	சாவான்	சாகும்
4	விடு	விட்டான்	விடுகிறான்	விடுவான்	விடும்
5	பெறு	பெற்றான்	பெறுகிறான்	பெறுவான்	பெறும்

II. ந்த்/கிற்/ வ் பிரிவு

இதில் 10 உட்பிரிவுகள் வரும்.

1. புகழ் - புகழ்ந்தான் - புகழ்கிறான் - புகழ்வான் - புகழும்

2. விழு – விழுந்தான் - விழுகிறான் - விழுவான் - விழும்

3. அறி – அறிந்தான் - அறிகிறான் - அறிவான் - அறியும்

4. நோ – நொந்தான் - நோகிறான் - நோவான் - நோகும்

5. வே – வெந்தான் - வேகிறான் - வேவான் - வேகும்

6. வா – வந்தான் - வருகிறான் - வருவான் - வரும்

7. அகல் - அகன்றான் - அகல்கிறான் - அகல்வான் - அகலும்

8. கொல் - கொன்றான் - கொல்கிறான் - கொல்வான் - கொல்லும்

9. ஆள் - ஆண்டான் - ஆள்கிறான் - ஆள்வான் - ஆளும்

10. கொள் - கொண்டான் - கொள்கிறான் - கொள்வான் - கொள்ளும்.

III. த்/கிற்/ப்/ பிரிவு

1. ஓடு – ஓடினான் - ஓடுகிறான் - ஓடுவான் - ஓடும்

2. சொல் - சொன்னான் - சொல்கிறான் - சொல்வான் - சொல்லும்

3. போகு – போனான் - போகிறான் - போவான் - போகும்.

IV. த்/கிற்/ப்/பிரிவு

1. கல் - கற்றான் - கற்கிறான் - கற்பான் - கற்கும்

2. கேள் - கேட்பான் - கேட்கிறான் - கேட்பான் - கேட்கும்

3. உண் - உண்டான் - உண்கிறான் - உண்பான் - உண்ணும்

4. பூண் - பூண்டான் - பூணுகிறான் - பூண்பான் - பூணும்

5. காண் - கண்டான் - காண்கிறான் - காண்பான் - காணும்

6. தின் - தின்பான் - தின்கிறான் - தின்பான் - தின்னும்.

V. ந்த்/கிற்/ப்/ பிரிவு

1. நில் - நின்றான் - நிற்கிறான் - நிற்பான் - நிற்கும்

VI. த்த்/க்கிற்/ப்ப்/ பிரிவு

1. படி – படித்தான் - படிக்கிறான் - படிப்பான் - படிக்கும்.

VII. ந்த்/ க்கிற்/ ப்ப்/ பிரிவு

1. நட – நடந்தான் - நடக்கிறான் - நடப்பான் - நடக்கும்

வினைகளை அவை ஏற்கும் இநற்தகால – நிகழ்கால – எதிர்கால ஓட்டுகளின் அடிப்படையில் ஏழு பெரும் பிரிவுகளாகவும் உருபனியல் விதிகளின் அடிப்படையில் மேலும்

சிறு பிரிவுகளாகப் பிரித்து சீரற்றுத் திரிபுறும் வினைகளையும் கருத்திற்கொண்டு 27 பிரிவுகளாகப் பிரிக்கலாம். அவையாவன:

செய் - பிரிவு, அழு – பிரிவு, சாகு – பிரிவு, விடு – பிரிவு, பெறு – பிரிவு, வே –பிரிவு, புகழ் - பிரிவு, நோ – பிரிவு, அறி – பிரிவு, அகல் - பிரிவு, கொல் - பிரிவு, ஆள் - பிரிவு, கொள் - பிரிவு, ஓடு – பிரிவு, சொல் - பிரிவு, போ – பிரிவு, கல் - பிரிவு, கேள் - பிரிவு, உண் - பிரிவு, பூண் - பிரிவு, காண் - பிரிவு, தின் - பிரிவு, நில் - பிரிவு, படி - பிரிவு, நட – பிரிவு.

முற்றுநிலை உருபனியல் பகுத்துக்குறித்தல்

(Finite State Morphological Parsing)

உள்ளீடு	**உருபனியல் பகுத்துக்குறித்தல் வெளியீடு**
கால்கள்	கால் + பெயர் + பன்மை
கால்	கால் + பெயர் + ஒருமை
கடாக்கள்	கடா + பெயர் + பன்மை
மரங்கள்	மரம் + பெயர் + பன்மை
கற்கள்	கல் + பெயர் + பன்மை
முட்கள்	முள் + பெயர் + பன்மை

முதல் குத்துநிலை வரிசையில் தரப்பட்ட வடிவங்களுக்கு இரண்டாவது குத்துநிலை வரிசையில் தரப்பட்டவாறு வெளியீட்டைப் பெறுதல் நமது நோக்கமாகும். இரண்டாவது வரிசையில் ஒவ்வொரு சொல்லின் பகுதியும் உருபனியல் பண்பு கூறுகளாகும். எடுத்துக்காட்டாக + பெயர் என்பது அந்த சொல் ஒரு பெயர் என்பதையும் + ஒருமை என்பது ஒருமை எண் என்பதையும் + பன்மை என்பது பன்மை எண் என்பதையும் குறிப்பிடும்.

உருபனியல் பகுத்துக்குறிப்பான் (Morphological Parser) உருவாக்குவதற்கு பின்வருவன தேவை.

1. அகராதி (Lexicon) பகுதிகள் மற்றும் ஒட்டுகள் இவற்றின் அடிப்படைச் செய்திகளுடன் கூடிய (அதாவது பத்தி பெயரா வினையா போன்ற செய்திகள்) அடங்கிய பட்டியல்

2. உருபன் அமைப்பொழுங்கு ஒரு சொல்லுக்குள்ள எந்த வகையைச் சார்ந்த உருபன்கள் பிற வகுப்புகளைத் தொடர்ந்து வரும் என்பதை விளக்குகின்றது. உருபன் வருகை முறையின் மாதிரி எடுத்துக்காட்டாக தமிழில் பன்மை உருபு பகுதியைத் தொடர்ந்து வரும். பகுதிக்கு முன்னர் வராது போன்ற செய்திகளைக் கூறலாம்.

மரம் + கள் +ஐ

- கள் - மரம் - ஐ
- ஐ – கள் - மரம்
- கள் - ஐ – மரம்
- மரம் - ஐ – கள்

3. சந்திவிதிகள்/உருபஒலியனியல் விதிகள் அல்லது எழுத்தலகு விதிகள் (Morphophonemic Rules Orthographic Rules) தமிழில் சந்தி விதிகள் சாதாரணமாய் இரு உருபன்கள் இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்ததப் பயன்படுகின்றன.

ஆங்கிலம் போன்ற மொழிகளில் எழுத்துக்கூட்டல் (Spellers) விதிகள் சாதாரணமாய் இரு உருபன்கள் இணைகையில் ஏற்படும் மாற்றங்களை மாதிரிப்படுத்தும்.

மரம் + கநளள் > மரங்கள்

மரம் + ஐ > மரத்தை

City + s > Cities

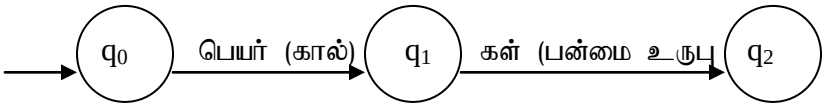
Box + s > Boxes)

உருபன் அமைப்பொழுங்கு பற்றிய அறிவை மாதிரிப்படுத்த எவ்வாறு முற்றுநிலைத் தானியங்கியைப் பயன்படுத்துவது உருபன்களைத் தெரிந்து கொள்ளலின் துணைச் சிக்கல்களை அகராதியில் எவ்வாறு எளிமையாக உருப்படுத்தம் செய்வது என்பது அடுத்து எதிர் கொள்ள வேண்டிய பிரச்சனைகள் ஆகும். இதற்கு நாம் முற்றுநிலை மாற்றிகள் (Finite State Transducers) அகராதியில் உருபனியல் பண்புக் கூறுகளை மாதிரிப்படுத்த அறிமுகப்படுத்த வேண்டி வரும். இறுதியில் எவ்வாறு எழுத்தலகு விதிகளை அல்லது உருபொலியனியல் விதிகளை மாதிரிப்படுத்த முற்றுநிலை மாறிகளை பயன்படுத்தலாம் என்பதையும் தர வேண்டும்.

அகராதியும் உருபன் அமைப்பொழுபங்கும் (Lexicon and Morphotactics)

அகராதி சொற்களை வைப்பதற்கான ஒரு களஞ்சியம் எளிய அகராதி மொழியின் எல்லா சொற்களையும் இயற்பெயர்களையும் கொண்டிருக்கும்.

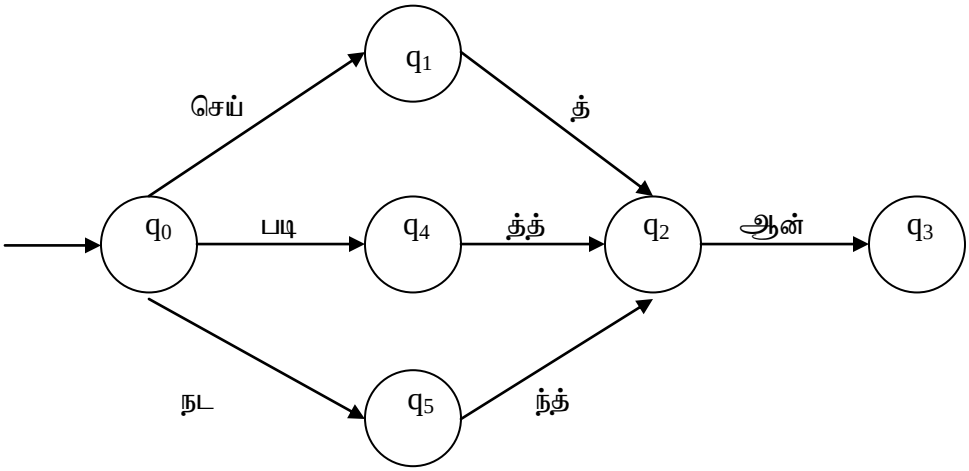
ஒரு மொழியில் உள்ள எல்லாச் சொற்களையும் பட்டியலிடுவது சௌகாரியமானதாகவோ இயலாமலோ இருப்பதால் கணினி அகராதி சாதாரணமாக மொழியின் ஒவ்வொரு பகுதியையும் ஒட்டுகளையும் அவை எவ்வாறு இணையும் என்ற செய்தியைத் தருகின்ற உருபன் அமைப்பொழுங்கின் உருப்படுத்தத்துடன் பட்டியலிட்டு கட்டமைப்பு செய்து தரும் உருபன் அமைப்பொழுங்கை பல வகைகளில் மாதிரிப்படுத்தலாம். மிகப் பொதுவான ஒரு மாதிரி முற்றுநிலைத் தானியங்கியாகும். பின் வருவது தமிழ்ப் பன்மை திரிபின் முற்றுநிலை மாதிரியாகும்.



பின்வருவது தமிழ் வினைத்திரிபின் முற்று நிலை மாதிரியாகும்

'செய்' – வகுப்பு வினைகள் படி – வகுப்பு வினைகள் செயல் - வகுப்பு வினைகள்

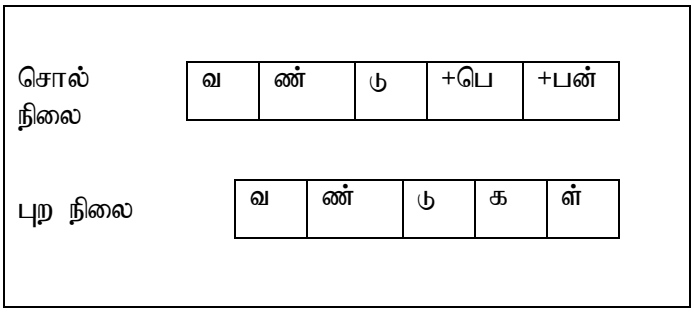
செய் வகுப்பு வினைகள்	படி – வகுப்பு வினைகள்	நட – வகுப்பு வினைகள்	செய் வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படி – வகுப்பு வினைகளின் இறந்தகால ஒட்டு	நட – வகுப்பு வினைகளின் இறந்தகால ஒட்டு	படர்க்கை ஆண்பால் ஒட்டு
செய் பெய்	படி பிடி குடி மற	நட அள கட	த்	த்த்	ந்த்	ஆன்



நாம் முற்றுநிலைத் தானியங்கியை பயன்படுத்தி உருபனியல் தெரிந்துகொள்வதின் சிக்கலை தீர்க்கலாம். அதாவது உள்ளீடு செய்கின்ற எழுத்து சரியான தமிழ்ச் சொல்லா இல்லையா என்று உறுதிசெய்து கொள்ளலாம்.

முற்றுநிலை மாறிகளின் உதவியால் உருபனியல் பகுத்துக்குறித்தல்

கொஸ்கெனிமி (Koskenniemi) எடுத்துச் சொன்ன இருநிலை உருபனியல் (Two Level Morphological) உருபனியல் பகுத்துக்குறித்தலைச் செய்யலாம். இருநிலை உருபனியல் ஒரு சொல்லை உருவாக்குகின்ற உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் சொல் நிலைக்கும் (Lexical Level) இறுதிச் சொல் வடிவின் உண்மையான எழுத்துக் கூட்டலை உருப்படுத்தம் செய்யும் புற அமைப்புக்கும் இடையே உள்ள பொருத்தத்தை உருப்படுத்தம் செய்கின்றது. எழுத்துக் கோர்வைகளால் ஆன மரங்கள் போன்ற புற அமைப்பு நிலையை மரம் + பெயர் + பன்மை என்ற உருபன் பண்புக்கூறுகளான தொடர்ச்சியின் சொல் நிலையாக பொருத்தம் செய்யும் விதிகளால் பொருத்தம் செய்வது உருபனியல் பகுத்துக் குறித்தலாகும்.



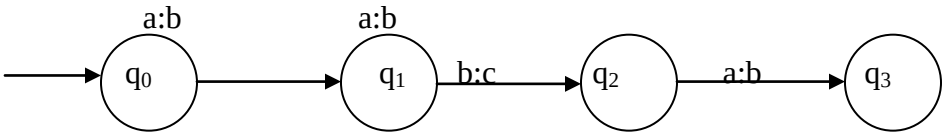
சொன்மை மற்றும் புற நாடாக்களின் எடுத்துக்காட்டு

இவ்வாறு இருநிலைக்களுக்கிடையே உருபடுத்தத்தைச் செயல்படுத்த நாம் பயன்படுத்தும் தானியங்கி முற்றுநிலை மாற்றியாகும் (FST) மாற்றிகள் ஒரு குமமக்குறியீடுகளுக்கும் மற்றொரு குழுமக்குறியீடுகளுக்கும் இடையே உருபடுத்தம் செய்கின்றது. முற்றுநிலை மாற்றி இதை முற்றுநிலைத் தானியங்கி வழியாகச் செய்கின்றது. இவ்வாறு நாம் ஒரு முற்றுநிலை மாற்றியை கோர்வைகளின் இணைகளை தெரிந்துக்கொள்கின்ற அல்லது உருவாக்குகின்ற இரு நாடா முற்றுநிலை தானியங்கியைக் காணலாம். இவ்வாறு முற்றுநிலை மாற்றிக்கு முற்றுநிலை தானியங்கியை விட கூடுதல் பொதுமைச் செயால்பாடு இருக்கின்றது. முற்றுநிலை தானியங்கி ஒரு முறையான மொழியை கோர்வைகளின் ஒரு குழுமமாக விவரணை செய்கின்றது. ஆனால் முற்றுநிலை மாற்றி ஒரு முறையான மொழியை கோர்வைகளின்

குழுமாங்களுக்கு இடையே உள்ள உறவாகக் விவரணை செய்கின்றது. இது முற்றுநிலை மாறியை ஒரு கோர்வையை உருவாக்கும் ஒரு பொறியாகக் காட்டுகின்றது.

ஒரு உள்ளீட்டாலும், வெளிட்டாலும் ஒவ்வொரு வில்லும் அடையாளப்படுத்தப்பட்ட முற்றுநிலைப்பொறி முற்றுநிலை மாற்றியாகும்.

ஒரு தொடர்பை முற்றுநிலைமாற்றியாக உருப்படுத்தம் செய்யலாம்.



இணைகளின் ஒரு முற்றுநிலை அல்லது எச்சநிலைக் குழுமம் ஒரு உறவாகும்.

R = (1,1), (2,4), (3,9), (4,16), (5,25)

S = (a, A), (b, B), (c, C), (d, D), (e, E). …

U = (கேள்வி i, பதில் i)

V = (ராமன் வீட்டுக்குப் போய்க்கொண்டிருக்கிறான் Raman is going home)

W = (a,b), (aaa,bbb), (aaaa,bbbb).....

ஒரு உறவை பார்வை அட்டவணை (look up table) (உள்ளீடு ,வெளியீடு) ஆக உருப்படுத்தம் செய்யலாம். அந்த அட்டவணை எல்லையற்ற அளவு பெரியதாய் இருக்கலாம்

பின் தரப்பட்டுள்ளது மாற்றிகளின் நாலுமடங்கு வழிகளின் சுருக்கமாகும்.

1. **அறிவானாக முற்றுநிலைமாற்றி (FST as recognizers)**

கோர்வைகளின் ஜோடியை/இணையை உள்ளீடாகவும் வெளியீடாகவும் எடுத்துக்கொண்டு கோர்வை இணைமொழியின் கோர்வை இணையாக இருந்தால். ஏற்றுக்கொள்ளவும் இல்லாவிட்டால் விட்டுவிடவும் செய்யும் மாற்றி முற்றுநிலைமாற்றி தெரிவானாகும்.

2.உருவாக்கியாக முற்றுநிலைமாற்றி (FST as generator)_

மொழியின் கோர்வைகளின் இணைகளை வெளியீடு செய்கிற பொறி முற்றுநிலை மாற்றி உருவாக்கியாகும் இவ்வாறு வெளியீடு "ஆம்" (அ) இல்லை என்பாதம் மற்றும் வெளியீட்டுக் கோர்வைகளின் இணையாகும்.

3. மொழிபெயர்ப்பியாக முற்றுநிலை மாற்றி (FST as translator)

ஒரு கோர்வையைப் படித்து மற்றொரு கோர்வையை வெளியீடு செய்கிற பொறி முற்றுநிலைமாற்றி மொழிபெயர்ப்பியாகும்.

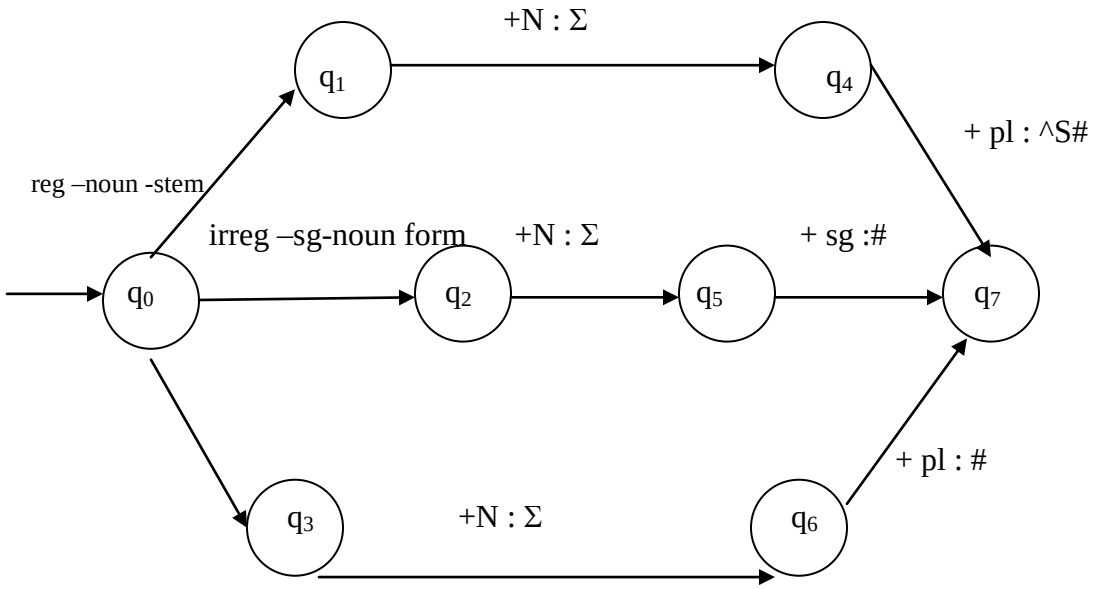
4. உறவுபடுத்தியாக முற்றுநிலைமாற்றி (FST as realtor)

குழுமங்களுக்கு இடையே உறவுகளைக் கணிக்கிற பொறி முற்றுநிலைமாற்றி உறவுப்படுத்தியாகும். எங்கு முற்றுநிலைத்தானியங்கிகள் முறையான மொழிகளுடன் சம உருவம் கொண்டிருக்கிறதோ அங்கு முற்றுநிலைமாற்றிகள் முறையான உறவுகளுடன் சமஉருவம் கொண்டிருக்கும். கோர்வைகளின் குழுமங்களாக இருக்கிற முறையான மொழிகளின் இயற்கையின் நீட்சியான கோர்வைகளின் இணைகளின் குழுமங்கள் முறையான உறவுகளாகும்.

எ.கா

ஆங்கில பன்மைத் திரிபு அகராதியில் கலவை நிலையில் அமையும்.

Reg – Noun	Irreg – pl – Noun	Irreg –Sg -Noun
Fox	go:co:ese	goose
Cat	Sheep	Sheep
Dog	Mo:iu:Σs: ce	Mouse



ஆங்கில பெயர்திரிபின் மாற்றி q_1,q_2 என்பன ஏற்றுக்கொள்ளும் நிலைகள் ஆகையால், சீரான/முறையான பெயர்கள் பன்மைப் பின்னொட்டு பெறவோ பெறாமலோ இருக்கலாம். ^ என்பது உருபன் - எல்லைக் குறியீடு: # என்பது சொல் - எல்லைக் குறியீடு

எ.கா.

சொல்நிலை

F	O	X	+N	+pl

இடைப்பட்ட நிலை	f	o	x	^	s	#

எழுத்தலகு விதிகளும் இறுதி நிலை மாற்றிகளும்

ஆங்கிலத்திற்கு உருபன் எல்லையில் எழுத்துக் கூட்டல் மாற்றங்கள் தேவை. அதற்காக நாம் எழுத்துக் கூட்டல் விதிகள் அல்லது எழுத்தலகு விதிகளை அறிமுகம் செய்ய வேண்டிவரும்.

நாம் எழுத்துக் கூட்டல் மாற்றங்களை எளிய உருபன்களின் இணைப்பின் உள்ளீட்டாகவும் கொஞ்சம் மாறுபட்ட உருபன்களின் இணைப்பின் வெளியீடாகவும் எடுத்துக்கொள்ளலாம். பின்வருவது மூன்று நிலைகளைக் காட்டுகின்றது. சொல் நிலை, இடைநிலை, மற்றும் புறநிலை

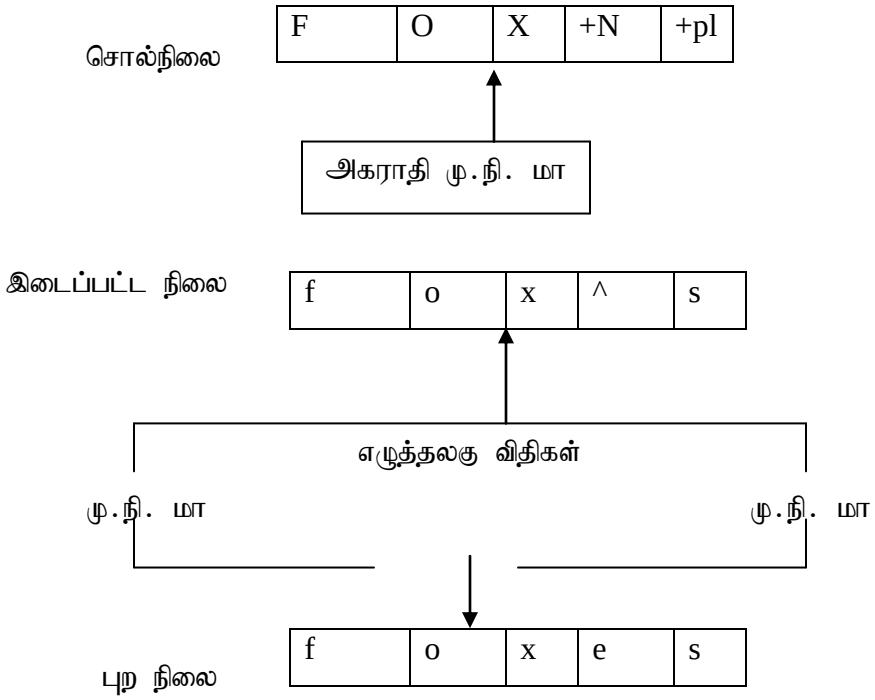
சொல்நிலை	F	O	X	+N	+pl

இடைப்பட்ட நிலை	f	o	x	^	s	#

புற நிலை	f	o	x	e	s	#

முற்றுநிலை மாற்றியின் அகராதியையும் விதிகளையும் சேர்த்தல் (Combining FST Lexicon and rules)

நாம் பகுத்துக்குறிக்கவும் உருவாக்கவும் அகராதியையும் விதி மாற்றிகளையும் இணைக்க தயாராகிவிட்டோம். பின்வரும் படம் பகுத்துக் குறிக்கவோ உருவாக்கவோ பயன்படும் இருநிலை உருபன் ஒழுங்குமுறையின் அமைவைக் காட்டுகின்றது.



அகராதி மாற்றி அதன் பகுதிகள் மற்றும் உருபன் பண்புக்கூறுகள் இவற்றைக் கொண்டு சொல் நிலைக்கும் உருபன்களை இணைவதை உருப்படுத்தம் செய்யும் இடைநிலைக்கும் இடையே பொருத்தம் செய்கின்றது. ஒவ்வொரு தனி எழுத்துக் கூட்டல் விதிக்கட்டுப்பாட்டையும் உருப்படுத்தம் செய்கின்ற ஒரு கூட்டம் மாற்றிகள் இடைநிலைக்கும் புறநிலைக்கும் இடையே பொருத்தம் செய்கின்றது.

பிசிகிமோ (Pckimmo)

கொசொகென்னிமி (Kosekenniemi) இன் இருநிலை உருபனியல் பரவலாகப் பயன்படுத்தப்படுகின்ற குறிப்பாக ஐரோப்பிய மொழிகளில் பயன்படுத்தப்படுகின்ற ஒன்று. பிசிகிமோ உருபனியல் ஆய்வியாக இரண்டு செயல்களைச் செய்ய இயலும்.

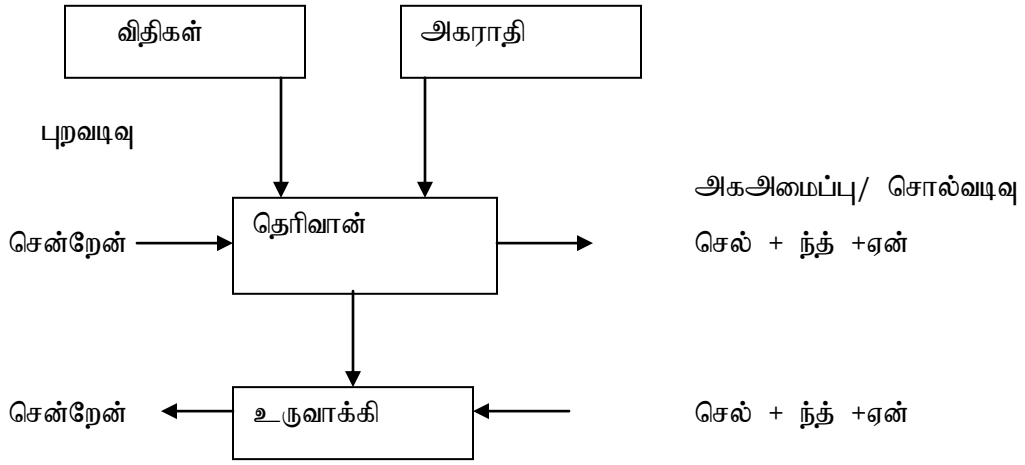
1. தெரிந்துக்கொள்ளுதல் (Recognition)
2. உருவாக்குதல் (Generation)

பிசிகிமோவின் உருவாக்கும் பகுதி வேரின் சொல்வடிவை உள்ளீடாக ஏற்கும்.

தெரிவான்/உணரி புற அமைப்பை உள்ளீடாக ஏற்கும் விதிகளை பயன்படுத்தும்: அகரதியைப் பார்க்கும் தொடர்புள்ள சொல்வடிவையும் பொருள் கோர்வையும் தரும். இதைப் பயன்படுத்துவதில் உள்ள எளிமையும் இம்மாதிரியின் சிக்கனம் அடிப்படையிலான முக்கியத்துவமும் என்னவென்றால் ஒரே குழும விதிகளை தெரிந்துகொள்ளும் செயல்பாட்டிற்கும் உருவாக்கும் செயல்பாட்டிற்கும் பயன்படுத்தலாம்.

விதிகளை இருதிசைகளிலும் பயன்படுத்தலாம். சில ஒலியன் விதிகளோ உருபன் விதிகளோ எழுத்தலகு விதிகளோ புற அமைப்பிலிருந்து, அக அமைப்பைத் தரும் மாற்றமாகக் கொள்ளப்பட்டால் அவற்றைத் திருப்பிச் செயல்பாட்டிற்காக (Reverse process) மீண்டும் எழுதத் தேவையில்லை.

திரும்புகையில் அது புற அமைப்பிலிருந்து அகஅமைப்பை ஆய்கிறது. இதைப் பின்வரும் படத்தில் காட்டலாம்.



இந்த மாதிரியில் பயன்படுத்தப்படும் இருநிலைகள் எல்லாம் மொழியியல் விவரணைகளாகும். இந்த விவரணைகள் மாற்றொலியன் பிறழ்ச்சிகள் (Allophonic alternations) உருப ஒலியன் பிறழ்ச்சிகள் (Morphophonemic alternations) அல்லது எழுத்தலகு வேறுபாடுகள் (Orthographic variations) இவற்றை கையாள இயலும்.

பிசிகிமோ இருநிலை மொழியியல் விவரணைகளை முற்றுநிலை அட்டவணைகளாக மாற்றும்.

விரிவாகக்கூறப் புகுந்தால் விதிக்கோப்பில் விதிகள் முற்றுநிலை மாற்றிகளாக (FST) உருபடுத்தம் செய்யப்படும். உருபன் அமைப்பொழுங்குக் கட்டுப்பாடு (Morphotatic Constraints) அகராதிக் கோப்பில் முற்றநிலைத் தானியங்கியாகக் (FSA) கூறப்படும்.

அகராதிப் பகுதி (Lexical Component)

பிசிகிமோவின் அகராதிப் பகுதி தெரிந்துக் கொள்ளும் செயல்பாட்டின் போது மட்டும்தான் நோக்கப்படும் உருவாக்கும் செயல்பாட்டில் நோக்கப்படமாட்டாது.

அகராதி ஒரு சொல்லின் உருப்படுத்தத்தைக் காட்டும் அதன் உருபன் அமைப்பொழுங்குக் கட்டுப்பாட்டை குறிப்பிடும். ஒவ்வொரு சொல்லிற்குப் பொருள் தரும் பிசிகிமோவின் ஒவ்வொரு அகராதியும் துணை அகராதிகளாகப் பகுக்கப்பட்டிருக்கும்.

ஒவ்வொரு துணை அகராதியும் உருபன் அமைப்பொழுங்கில் ஒரே மாதிரி செயலாற்றுகிற ஒரு வகுப்பைச் சார்ந்த சொற்களை உருப்படுத்தம் செய்யும்.

ஒவ்வொரு சொற்பதிவும் (Lexical Entry) ஒரு சொல்லையும் அதன் தொடர்ச்சி வகுப்பையும் அதன் பொருளையும் கொண்டிருக்கும்.

தொடர்ச்சி வகுப்பு அதைத் தொடர்ந்து வரும் வேறுபட்ட உருபன்களையும் இலக்கணத் தன்மையின் உருபன் அமைப்பொழுங்கு விதிகள் எங்கு கூறப்பட வேண்டும் என்பதைக் காட்டும்.

ஒரு சொல்லைத் தொடர்ந்து வரும் துணை அகராதிக் குழுவை விவரிக்கும் பிறழ்ச்சியின் பட்டியல் இருக்கும்.

எடுத்துக்காட்டாக,

கால் என்ற சொல்லை எடுத்துக் கொள்வோம் அது –ஐ- என்ற இரண்டாம் வேற்றுமை உருபையும் உக்கு என்ற நான்காம் வேற்றுமை உருபையும் இல் என்ற இட வேற்றுமை உருருபையும் இலிருந்து என்ற நீக்கல் வேற்றுமை உருபையும் கள் என்ற பன்மை உருபையும் ஆ என்ற வினா பின்னொட்டையும் ஏற்கும். சேய்மை கட்டு அடையில் அ என்பது அந்த என்பதன் மாற்று உருபாகும்.

இயல் 4

தொடர்களை ஆய்தல்

கணினி மொழி ஆய்விற்கான இலக்கண வடிவமைப்புகள்

இலக்கண வடிவமைப்பு (Grammar Formalism)

ஒரு கூற்றின் பொருத்தமான பொருளைத் தருமாறு சொற்களுக்கு இடையே உள்ள உறவுகளை பிரித்தெடுக்கும் நுட்பத்தைத் தருவது இந்த இலக்கண வடிவமைப்பின் முக்கியமான செயல்பாடாகும்.

முறையான இலக்கணம் (Formal Grammar)

வாக்கியங்களை பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது அந்த வாக்கியங்களுக்கு அமைப்மபைத் தருகின்றது. மொழியைப்

பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல்.

சூழல் வரையறை இல்லாத இலக்கணம் (Context Free Grammar)

சூழல் வரையறை இல்லாத இலக்கணம் மரபு மொழியியலரால் அண்மை உறுப்பு இலக்கணங்கள் என்றும் வழியமைப்பு மொழித் திட்டமிடுபவர்களால் பேக்ஸ் இயல்பான வடிவு (Bgackes Normal Form) என்றும் சில கணினி பயன்பாட்டில் மறுதரவு அமைப்பொழுங்கு (Recursive Pattern) என்றும் அழைக்கப்படுகின்றது. இவைகளைப் பொதுவாகத் தொடரமைப்பு இலக்கணம் (Pharse Structare Grammar) என்பதன் குறிப்பிட்ட வகை எனலாம்; இவைதான் ஆக்கமுறை இலக்கணத்திதற்கு (Generative Grammar) அடிப்படையில் அமைந்தது.

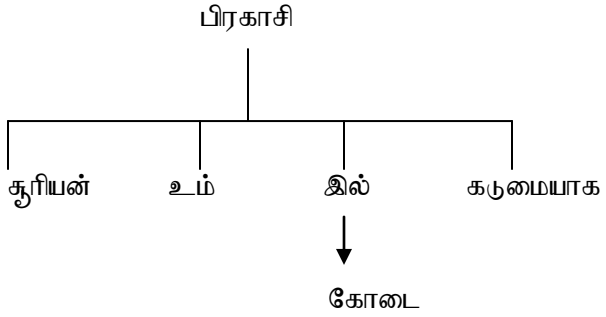
அமைப்புக்கான முன்று அணுகுமுறைகள்

புடைப்பெயர்வு பின்னல் (Transition Network) தட்டையான அமைப்பு (Foot Structure) உடையவை நம்முடைய உள்ளுணர்வு உறுப்பமைவை எதிர்பார்க்கின்றது. இதை வாக்கியம் என்பது பின்னலால் விற்களின் தொடர்ச்சியாகப் பொருந்துவதாக கொள்ள இயலாது. வாக்கியம் துண்டுகளால் அல்லது தொடர்களால் ஆனது. தொடர்களின் அமைப்பு பொருள் எவ்வாறு பெறப்படுகின்றது என்று கூறுவதால் இவ்வகையிலான அமைப்பாக்கம் முக்கியமானதாகும். மொழி அமைப்பை விளக்க பல வகைப்பட்ட விதிகளின் வடிவமைப்புகள் (Parmulations) இருக்கின்றன. எவ்வாறு மொழி அமைப்பை நன்றாக விளக்கலாம் என்ற அணுகுமுறையின் அடிப்படையில் அவற்றின் விளக்கங்களே வேறுபடுகின்றன.

தலை மற்றும் அடை அணுகுமுறை

இலக்கணத்தின் மரபுத்தன்மையிலான பல அறிமுகங்கங்கள் அமைப்புகளை அடிப்படை அமைப்பொழுங்கிலிருந்து நீட்சி செய்யப்பட்ட அடை செய்யப்பட்ட அமைப்பாக விவரணை செய்கின்றனர். எடுத்துக்காட்டாக, வாக்கியத்தின் அக அமைப்புப் பெயர்த் தொடரைத் தொடர்ந்து வரும் வினை என்று கொள்ளலாம். கூடுதலான சொற்களில் பெயர் அடைகள் சுட்டடைகள் (அடைகொளி அடைகள்) என்பன பெயரை அடை செய்வதாகவும் வினையடை போன்றவை வினையை அடை செய்வதாகவும் அடைகளையும் அடைகளால் விவரிக்கலாம். ஒவ்வொறு நிலையிலும் தனிச்சொல்லாக ஒரு தலை இருக்கும். அதற்கு ஒன்றோ அதற்கு மேலோ அடைகள் வரலாம்.

எடுத்துக்காட்டாக, சூரியன் கோடையில் கடுமையாகப் பிரகாசிக்கும் என்ற வாக்கியத்தைப் பின்வருமாறு ஆயலாம்.



பொருளையும் தொடரியலையும் உள்ளடக்கிய காரணத்தால் தலை அடை அணுகுமுறை மொழியியலில் முக்கிய இடம் வகிக்கின்றது.

பெயர் பொருள் (சாதனம்) எப்படி இருக்கிறது என்று கூறும் பெயரடை கூடுதலான விவரணைத் தரும் வினை என்ன நடந்தது என்று கூறும் வினையடை முறை காலம், இடம் மற்றும் பிற செய்திகளைத் தரும் இவ்வகையிலான சொல் அமைப்பிற்கும் கருத்துரு அமைப்பிற்கும் இடையிலான பொருத்தம் மொழியைப் பொருளை வெளிப்படுத்துகிற ஒருமித்த ஒழுங்குமுறையாகப் பார்ப்பதால் பயனுள்ளதாய் இருக்கின்றது.

தலை–அடை இலக்கணம் சார்பு இலக்கணமாக வடிவமைக்கப்பட்டு பயன்படுத்தப்படுகின்றது. சார்பு இலக்கணக் கொள்கை எவ்வாறு அமைப்பொழுங்கின் நிலையான அறிவு அமைப்பாக்கம் செய்யப்படுகின்றது. அல்லது வாக்கியம் ஆய்வு செய்யப்படுகின்றது என்று கூறாமல் அமைப்புகளை விளக்குவதின் வழிகளை வலியுறுத்துகின்றன.

அண்மை உறுப்பு அணுகுமுறை

தொடரமைப்புச் சட்டகம் (Phrase Structure Formwork)

1. அண்மை உறுப்பு பகுநிலை (Analytic) அடிப்படையிலானது.
2. அடிப்படை மொழி அலகுகளைக் காண்பதற்காக அறிவியல் கண்டுபிடிப்புச் செயல்முறை (Discovery Procedures) செயல்பாட்டால் உருவானது.
3. வாக்கியங்களை இரண்டு அண்மை உறுப்புப் பாகங்களாகவும் ஒவ்வொரு சிறு பாகத்தையும் மேலும் சிறு அலகுகளாகவும் (உருபன் வகை) பகுப்பது இக்கோட்பாட்டின் கொள்கையாகும்.
4. சில மொழியியலார் வாக்கியத்தை பெரிய அலகுகளாகக் கொண்டு மேலிருந்து கீழ் நோக்கி (Top- Down) ஆய்வாளர்கள் சிலர் கீழிருந்து ஆய்வார்கள்.
5. அண்மை உறுப்பு ஆய்வு அடையாளப்படுத்தப்பட்டால் மரபிலக்கணப் பகுத்துக் குறித்தல் (Parsing) போலவே அமையும். அதாவது வாக்கியத்தை இலக்கண வகைப்படுகளாக பகுத்தல் நிகழும்.

6. மரபு அடிப்படையிலான வாக்கியம் பகுப்பாயவும் பகுத்துக் குறித்தலும் (Parsing) தொடரமைப்பு இலக்கணம் எனப்படலாம். இவ்விலக்கணம் பின்வரும் கேள்விகளுக்கு விடை தருகின்றது.
    1. ஒரு வாக்கியத்தின் உருப்புகள் யாவை?
    2. அவை எவ்வாறு ஒழுங்குபடுத்தப்பட்டுள்ளன
7. மரபிலக்கணங்கள் வாக்கியங்களையும் தொடர்களையும் விளக்கப் பின்வரும் கூற்றுகளைத் தருகின்றன.
    1. ஒரு வாக்கியத்தில் பெயர்த்தொடரும் (Noun Phrase) வினைத்தொடரும் அடங்கும்.
    2. ஒரு பெயர்த்தொடரில் ஒரு பெயர் மட்டுமோ அடைகொளி அடை + பெயரோ அடைகொளி அடை + பெயரடை + பெயரோ வரும். எடுத்துக்காட்டாக, 1. அந்தப் பையன் 2. அந்த நல்லப் பையன்.
    3. ஒரு வினைத்தொடரில் பெயர் மட்டுமோ வினை மட்டுமோ வினை + துணை வினையோ பின்வருபுத் தொடர் + வேற்றுமைத் தொடர் + வினையோ வரும்.

       எடுத்துக்காட்டு,

       வந்தான்

       வந்துக்கொண்டிருந்தான்

       வீட்டிலிருந்து வந்துக்கொண்டிருந்தான்

       பழத்தைச் சாப்பிட்டான்
    4. பின்னுருபுத் தொடரில் வேற்றுமைத் தொடரும் பின்னுருபும் வரும்,
       எடுத்துக்காட்டு, *வீட்டிலிருந்து*
    5. வேற்றுமைத் தொடரில் பெயர்த்தொடரும் வேற்றுமை உருபும் வரும்.
       எடுத்துக்காட்டு, *வீட்டை, வீட்டின்*

திட்டவட்டமற்ற சிக்கனமில்லாத இம்மாதிரியான விளக்கக்கூறுகள் தருவதற்குப்பதிலாக சில மரபு மற்றும் கருத்துச் சாயல் அடிப்படையில் ஒரு குழு ஆக்கவிதிகள் (Generation rules) அல்லது விரித்தெழுது விதிகள் (Rewrite riles) தரலாம்.

அம்மாதிரியான விதிகளை தொடரமைப்பு விதிகள் (Phrase Structure Rules) எனலாம். தொடரமைப்பு விதிகள் பின் மாதிரியான பொது விதிச் சுருக்கத்தை (Rules-schema) கொண்டிருக்கும்.

அ__ஆ

அதாவது 'அ' என்ற குறியீட்டைத் தொடக்ககுறியீடு என்ற குறியீடாக விரித்தெழுதலாம்.

விரித்தெழுதும் செயல்பாட்டைத் தொடங்கிவைப்பதற்குப் பயன்படுத்தும் குறியீட்டைத் தொடக்ககுறியீடு எனலாம்.

எ.கா

வா- பெ.தொ + வி.தொ.

என்ற விரித்தெழுது விதியில் 'வா' என்பது தொடக்கக்குறியீடு மேற்சொன்ன விதி ஒரு வாக்கியத்தில் பெயர்த்தொடரும் வினைத்தொடரும் வரும் என்று தெரிவிக்கின்றது.' +' என்ற குறி பிணைப்பைக் (Concatenation) குறிப்பிடப் பயன்படுகின்றது.

பெ.தொ, வி.தொ என்ற புலக்குறிகளை (labels) விரித்தெழுதவோ செய்யலாம்.

எ.கா

பெ.தொ – அடைகொளி அடை + பெயர்

என்ற விதி பெ.தொ கணுவை விரித்தெழுதுகின்றது.

(அடைகொளி அடையில் அந்த, இந்த, என், உன், அவனது, அவளது, அதன், அவற்றின் என்பன அடங்கும்).

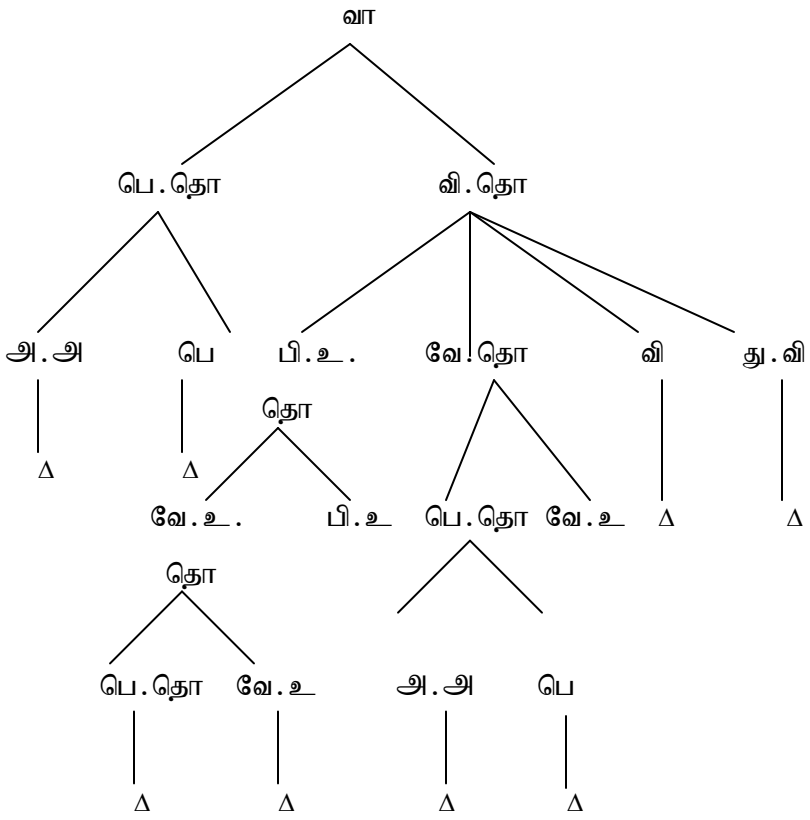
பையன், சிறுமி போன்றன பெயர்ச்சொல்லில் அடங்கும். கணு என்பது கிளைப்படத்தில் ஒன்றோ அல்லது அதற்கு மேற்பட்டகிளைகளோ வெளிவரும் தொடக்கவிடமாகும். தமிழில் வி.தொ என்பதை பி.உ.தொ + வே.தொ + வி.தொ + து.வி என விரித்து எழுதலாம்.

வி.தொ – பி.உ.தொ + வே.தொ + வி + து.வி

பி.உ.தொ என்ற கணுவைப் பின்வருமாறு விரித்து எழுதலாம்.

பி.உ.தொ – பெ.தொ + பி.உ

இதுவரை தந்த விதிகளைப் பின்வருமாறு கிளைபடமாகத் தரலாம்.



மேற்கண்ட கருத்துச் சாயல்களைப் பயன்படுத்தி தமிழ் இலக்கணத்தில் ஒரு பகுதிக்கு தொடரமைப்பு விதிகள் உருவாக்கலாம்.

வா – பெ.தொ + வி.தொ

வி.தொ – வே.தொ + (பி.உ.தொ + வி +து.வி

து.வி _ கொண்டிரு, இரு, வேண்டு, பார், போ

வி - கொல், வெட்டு, நட, பார், ஓடி

பெ.தொ – அ.அ + (பெ.அ) + பெ

பெ _ பையன், பாம்பு, மரம், மேஜை, சிறுமி, பெண்.

அ.அ - அந்த, ஒரு, ஓர், தன்

பெ.அ - வயதான, இளமையான, அழகான

பி.உ.தொ – வே.தொ + பி.உ

வே.தொ _ பெ.தொ + வே.உ

பி.உ _ பக்கத்தில், மேலே, கீழே, நோக்கி, அருகில், அடியில்

வே.உ _ ஐ, இல், .ஆல், உக்கு, ஓடு

மேற்கண்ட தொடரமைப்பு இலக்கணம் (Phrase sturcture grammar) பின் வரும் தமிழ் வாக்கியங்களை உருவாக்கும்.

அந்தப் பையன் ஒரு பாம்பைப் கொல்லப் பார்த்தான்

(பெ,தொ + வே.தொ + வி + து.வி)

அந்த அழகான பெண் தன் வீட்டை நோக்கி நடந்தாள்

(பெ,தொ + பி.உ.தொ + வி)

அந்த வயதான மனிதர் மரத்தை வெட்டப்போனார்.

(பெ.தொ + வே.தொ + வி + து.வி)

ஆனால் மேற்சொன்ன விதிகள் பின்வரும் பொருளற்ற வாக்கியங்களையும் உருவாக்கும்.

- மேஜை அந்தப் பெண்ணை மரத்திற்கு அருகில் கொன்றது.
- மரம் அந்தப் பையனைப் போகும்
- அந்தச் சிறுவன் அந்தச் சிறுமியை ஒடிக்க முடியும்
- மேஜை அந்த வயதான மனிதரைப் பார்க்கும்

மேற்சொன்ன விதிகள் சுழல் வரையறையில்லாத விதிகளாகும். அவற்றைச் சுழல் கட்டுண்ட விதிகளாக (Contex sensitive rules) மாற்றினால் மேற்சொன்ன வாக்கியங்களின் உருவாக்கத்தைத் தடை செய்யலாம்.

சூழல் வலையரை இல்லாத இலக்கணத்தைப் பயன்படுத்தி பகுத்துக்குறித்தல்

உள்ளீடு செய்யப்படும் கோர்வையைத் (sin) தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன இரண்டும் இணைந்த செயல்பாடுதான் பகுத்துக்குறித்தலாகும்.

வாக்கியத்தை தெரிந்து கொள்வதும் அதற்கு அமைப்பைத் தருகிறதும் ஆன செயல்பாடு தொடரியல் பகுத்துக்குறித்தல் எனப்படும்.

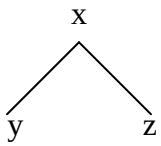
இவ்வாறு பகுத்துக்குறிப்பான் இரு செயல்களைச் செய்ய வேண்டும். ஒன்று ஒரு கோர்வைத் தரப்படும் போது அது பகுத்துக்குறிக்க வேண்டிய மொழியின் வாக்கியம் என்பதைத் தெரிந்து கொள்ள வேண்டும்.

இந்நிலையில் பகுத்துக்குறிப்பானுக்கு உள்ளடங்கிய தெரிந்துகொள்வான்கள் (built –in recognigeras) இருக்க வேண்டும்.

இரண்டாவது அது வெளியீடு செய்கிற அமைப்பை வாக்கியத்திற்கு தர வேண்டும்.

இலக்கணத்தில் இருக்கிற மொழியியல் செய்திகளை பகுத்துக்குறிப்பான்கள் சார்ந்திருக்க வேண்டும் என்பது இதன் மூலம் தெரிகின்றது. பகுத்துக்குறிப்பான்கள் மாறிகள் (transducors) என்ற பொருள் வகையைச் சார்ந்தது அதாவது வெளியீட்டு வசதிகளுடன் கூடிய தெரிந்துகொள்வான்களைப் (recogfnigers) பொறுத்த வரையில் இலக்கணங்களும்

பகுத்துக்குறிப்பான்களும் பண்பு வாய்ந்தவை x--→y என்ற விதி ஒரு உருவாக்கமாகத் தரப்பட்ட கலவைநிலை கட்டளையாகக் கருதப்பட வேண்டும்.



பகுத்துக்குறிப்பான் பெரும்பாலும் குறியீடுகளின் கோர்வையை ஏற்கும்: அதற்கு ஒரு விதிகையைப் பயன்படுத்தும் அது ஒரு கோர்வையின் விரித்தெழுது விதி போல் வரும்.

எடுத்துக்காட்டாக b---→d என்ற விதியைப் பயன்படுத்தி abc – என்பதை adc என்று விரித்தெழுதலாம் adc என்பதை adc என்று விரித்தெழுதலாம். 'ABC' 'ADC', மற்றும் 'Adc' என்பன வாக்கிய வடிவுகள் (Sententical forms) அல்லது ஆக்கங்கள் எனப்படும்.

'ADC' என்ற கோர்வை 'ABC' யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC ஆனது ABC யிலிருந்து நேரடியாகப் பெறப்பட்டது ஆகும்.

'ADC' என்ற கோர்வை 'ABC' யில் இருந்து ஒரு விதியைப் பயன்படுத்தி பெறப்படுவதால் ADC ஆனது ABC யிலிருந்து மறைமுகமாகப் பெறப்பட்டது ஆகும்.

ஒவ்வொரு படியிலும் பகுத்துக்குறிப்பான்கள் ஏதோ ஒரு அமைப்பை வெளியிடும்.

ஒரு குழும விதிகளின்படி வரக்கூடிய எல்லா அமைப்புகளையும் நாம் தெரிந்தால் ஒருவாக்கியத்தைப் பகுத்துக் குறிக்கலாம்.

பகுத்துக் குறிக்கும் நடைமுறைத்திறன் (parsing strategies):

மொழியியல் அறிவை பிரதிபலிக்கின்ற விதிகளின் அடிப்படையில் பகுத்துக்குறிப்பான் செயல்படம். வாக்கியங்களுக்கு அமைப்புத் தருகையில் ஒரு குழும விதிகளை பல நிரல்களில் நிறைவேற்றலாம்.

ஒவ்வொரு வேறுபட்ட நிரலும் வேறுபட்ட பகத்துக்குறித்தல் நடைமுறைத் திறத்துடன் தொடர்புடையது. பகுத்துக்குறிப்பான்களை அவற்றோடு தொடர்புடைய நடைமுறைத்திறன் அடிப்படையில் வகைப்பாடு செய்யலாம்.

பகுத்துக் குறித்தல் நடைமுறைத் திறனாக இரண்டு அளவீடுகள் அடிக்கடி பயன்படுத்தப்படுகின்றனவும் நிலைபேறு பெற்றவனவும் ஆகும்.

முதலாவது அது பகுத்துக்குறிக்கும் கோர்வைக்கு வெளியீடுகளின் மொழியியல் அமைப்பைக் கவனக் குவிப்பு செய்கிறது.

உள்ளீட்டுக் கோர்வையிலிருந்து தொடங்கி அமைப்பு கட்டப்படுவதனால் அந்த பகுத்துக்குறிப்பான் கீழிருந்து மேல் (bottom –up) செயல்படுவதாகக் கூறலாம். அல்லது

தொடக்கக் குறியீட்டிலிருந்து (கிளை அமைப்பின் வேர் அல்லது) தொடங்குவதனால் பகுத்துக்குயறிப்பான் மேலிருந்து கீழ் (top – town) செயல்படுவதாகக் கூறலாம்.

பகுத்துக்குறிக்கும் நடைமுறைத் திறன்களை வகைப்படுத்தும் அளவீடுகளை ஒரு எடுத்துக்காட்டு மூலம் விளக்கலாம். பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக்கொள்வோம்.

1. a. s--→ AB

   b. s--→ cd

   C. A--→ a

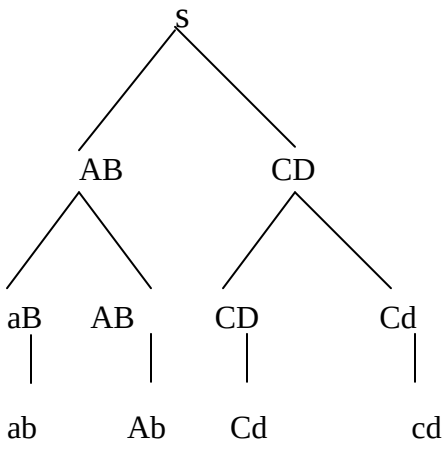
   d. B--→ b

   e. C---→ C

   f. D--→d

தரப்பட்டுள்ள ஒரு குழும விதிகளிலிருந்து அவற்றால் பெறப்படும் எல்லா ஆக்கங்களின் திட்டத்தைக் கட்ட முடியும். தொடக்க குறியீடான S- இல் இருந்து இரண்டு விதிகளைச் செயல்படுத்தலாம். விதி 1.a A,B இன் ஆக்கத்தையும் விதி 1.b. CD இன் ஆக்கத்தையும் விளைவிக்கும். A.B யில் விதி 1.c ஐயும் விதி 1.d ஐயும் பயன்படுத்தி முறையே a,b மற்றும் A.b என்ற கோர்வைகளைப் பெறலாம். இந்த எண்ணத்தைப் பின்பற்றி தரப்பட்ட ஒரு குழும விரித்தெழுது விதிகள் அனுமதிக்கம் ஆக்கங்கங்களின் எல்லா கோர்வைகளையும் கண்டுகொள்ளலாம் விளைவு பெரும்பாலும் கிளை வடிவாக உருப்படுத்தம் செய்யப்படும்.

2.



மேலே தரப்பட்ட கிளை அமைப்பு 1 இல் தரப்பட்ட விதிகளால் விளையும் ஆக்கங்களின் எல்லா கோர்வைகளையும் காட்டுகின்றது. கிளையில் வரும் ஒவ்வொரு

கனுவும் (A,B,C,D, aB, Ab, CD.cd.(d) வாக்கிய வடிவில் நடைமுறைப்படுத்தக் கூடிய வேறுபட்ட விதிகளின் தேர்வை உருப்படுத்தம் செய்யும்.

கிளை அமைப்பின் கிளைகள் (a,b, cd) அதாவது அடியில் வரும் கணுக்கள் மேலும் விதிகளின் பயன்படுத்த இயலாத வாக்கிய வடிவுகளை உருப்படுத்தம் செய்யும் இந்த கிளை அமைப்பு மொழியியலார் அவ்வாறு ஒரு வாக்கியத்தின் பாகங்கள் இணைகின்றன என்பதைக் காட்டுகின்ற மொழியியல் அமைப்பை உருப்படுத்தம் செய்யப் பயன்படுத்துகின்ற கிளையுடன் ஒப்பிட்டு குழப்பிக்கொள்ளக்கூடாது.

மேல் சொன்ன கிளை அமைப்பைக் காட்ட இயலும் ஆக்கங்களின் எல்லா கோர்வைகளையும் தருகின்றன.

இந்த வகையான கிளை அமைப்பின் அடிப்படையில் ஆழம் முதல் (னநிவா கசைளவ) அல்லது அகலம் முதல் (breadth – first) என்ற பகுத்துக் குறிக்கும் நடைமுறைத்திறன்களின் பாகுபாடு அமையும்.

மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்துக்குறித்தல் (Top – Down) Versus bottom –up Parsing)

பின்வரும் இரு எடுத்துக்காட்டுகள் எவ்வாறு மேலிருந்து கீழ் மற்றும் கீழிருந்து மேல் பகுத்துக்குறிப்பான்கள் பின்வரும் வாக்கியத்திற்கு அமைப்பு தருகின்றது என்பதை விளக்கும்.

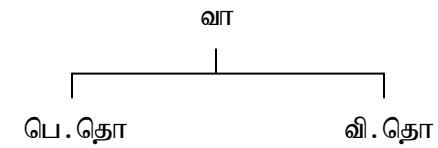
1. ஒரு பூனை எலியைப் பிடித்தது.

இரண்டும் பின்வரும் விரித்தெழுதும் விதிகளைப் பயன் படுத்தப்படுவதாக வைத்துக்கொள்வோம்.

2. a. வா ⟶ பெ.தொ + வி.தொ
2. b. பெ.தொ ⟶ அ.அ + பெ
2. c. வி.தொ ⟶ வே.தொ + வி
2. d. வி.தொ ⟶ வி
2. e. வே.தொ ⟶ பெ.தொ + வே உ
2. f. வி ⟶ பிடித்தது
2. g. பெ ⟶ பூனை
2. h. பெ ⟶ எலி
2. i. அ.அ ⟶ ஒரு
2. j. வே.உ ⟶ ஐ

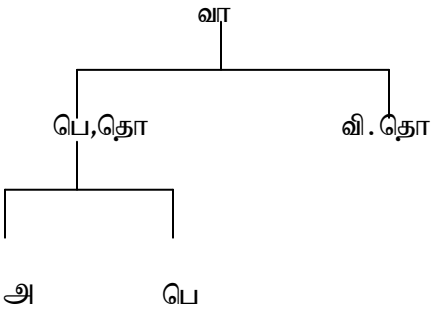
மேலிருந்து கீழ் பகுத்துக்குறித்தல் (மேல் -கீழ் பகுத்துக்குறித்தல்) (Top- Down parsing)

மேலிருந்து கீழ் பகுத்துக்குறிப்பான்கள் எப்பெழுதும் தொடக்கக் குறியீடான 'வா' என்பதிலிருந்து தொடங்கும் அதில் பயன்படுத்த விதிகளைக் கண்டு கொள்ளும் அதை விரிக்கும் எடுத்துக்காட்டில் 2.a என்ற விதிதான் இருக்கிறது. அதைப்பயன்படுத்தி விளையும் விளைவு பின்வரும் அமைப்பைத் தரும்.



இரு புதிய கணுக்கள் தோன்றும்.

இந்த கணுக்கள் இறுதிக் குறியீடுகளா (Forminal Symbols) என்று முதலில் பகுத்துக்குறிப்பான் பார்க்கும். அப்படியானால் அந்தக் குறியீடுகள் பகுத்துக் குறிக்கப்படுகின்ற கோர்வையுடன் பொருத்திப்பார்க்கப்படும். இல்லாவிடில் பகுத்துக்குறிப்பான் முதல் இறுதியுறா குறியீட்டுக் கணுவை வரிக்கும். மேற்கண்ட எடுத்துக்காட்டில் பெ. தொ வில் 2.b என்ற விதி பயன்படுத்தப்பட்டு பின்வரும் அமைப்பு விளையும்.



புதிதாக உருவாக்கப்பட்ட இந்தக் கணுக்களும் இறுதிக் கணுக்கள் அல்ல.

கீழிருந்து மேல் பகுத்துக்குறித்தல் (Bottam –up parsing)

கீழ் மேல் பகுத்துக்குறித்தல் உள்ளீடு செய்யப்பட்ட கோர்வையைத் தொடக்கமாகக் கொண்டு செயல்படத் தொடங்கி அதை வேறான 'வா' என்று சுருக்கும். அது ஒரு வாக்கியத்தை உள்ளீடாக ஏற்கும் சொற்களை (இறுதியுற்ற குறியிடுகளை) அவற்றின் வகைப்பாட்டால் இடம்பெயர்க்கும்.

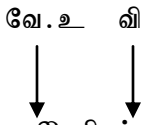
இதற்காக அது விரித்தெழுது விதிகளின் வலப்பக்கம் பார்த்து அவற்றை இடப்பக்கம் உள்ள வகைப்பாடாக சுருக்க வேண்டும்.

மீண்டும் ஒரு பூனை எலியைப் பிடித்தது என்ற வாக்கியம் 2a- j என்ற விதிகளால் கூறப்பட்டுள்ள இலக்கணத்தால் அமைப்பு பெறும்.

எடுத்துக்காட்டாக முதலில் 2f என்ற விதியைப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்.

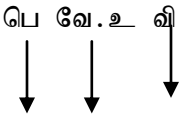
ஒரு பூனை ஒரு எலி ஐ பிடித்தது

2j. ஐப் பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்



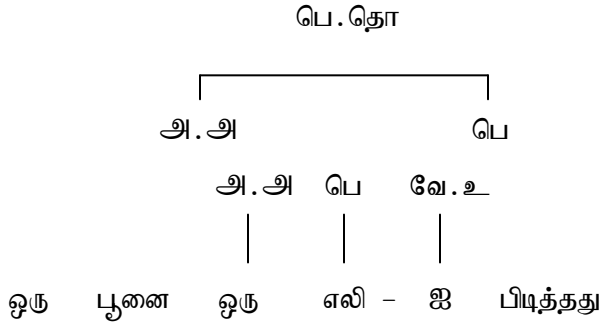
ஒரு பூனை ஒரு எலி - ஐ பிடித்தது

2h ஐ பயன்படுத்தி பின்வரும் அமைப்பு கிடைக்கும்

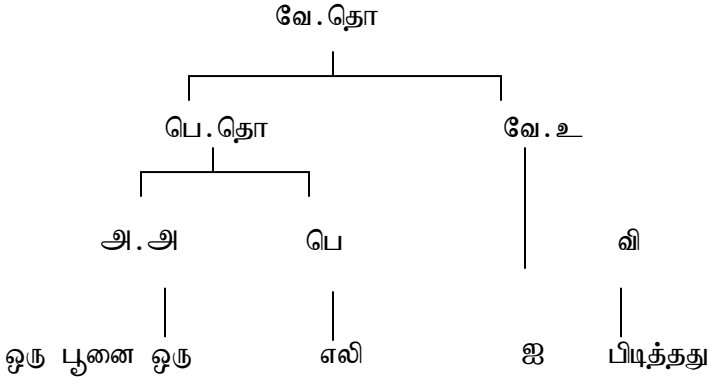


ஒரு பூனை ஒரு எலி ஐ பிடித்தது

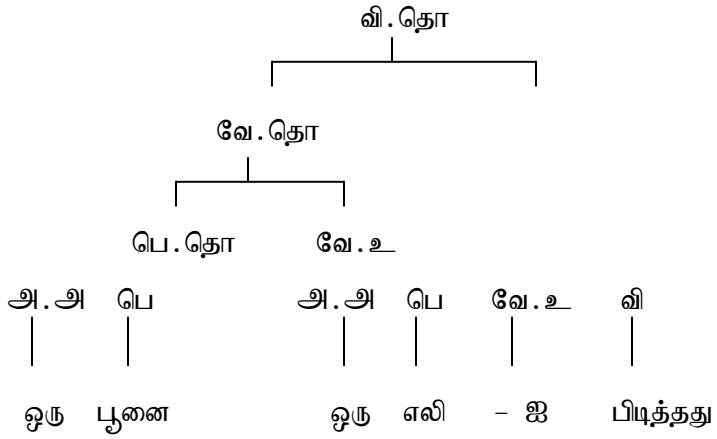
இந்த இடத்தில் 2b என்ற விதி 'அ.அ' என்பதையும் 'பெ' என்பதையும் இணைந்து அவற்றை பெ.தொ எனக் குறிக்கும்.



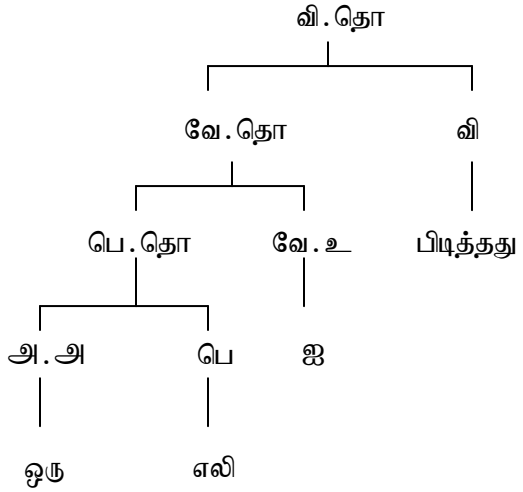
இந்த இடத்தில் 2e என்ற விதி வே.உ என்பதையும் பெ.தொ என்பதையும் இணைத்து அவற்றை வே.தொ என சுருக்கும்.



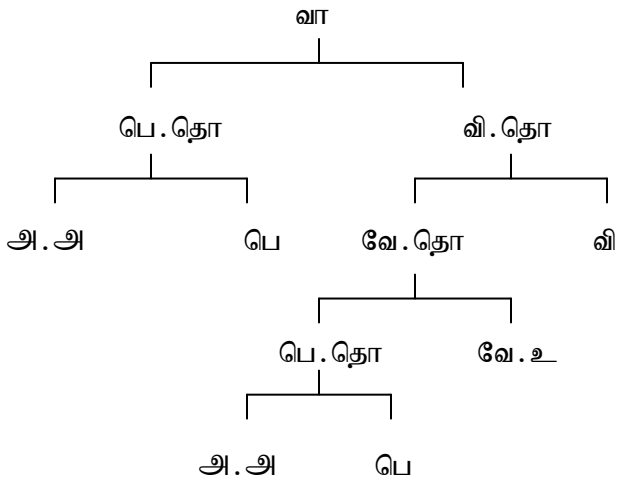
விதி 2c ஐ வே.தொ ஐயும் வி - ஐயும் இணைத்து வி.தொ எனச் சுருக்கும்.



விதி 2b அ.அ ஐயும் இணைத்து பெ ஆக சுருக்கும்.



விதி 2a பெ.தொ ஐயும் வி.தொ ஐயும் இணைத்து வா ஆகச் சுருக்கும்.



அ.அ பெ அ.அ பெ

| | | |

ஒரு பூனை ஒரு எலி - ஐ பிடித்தது

பகுத்துக் குறிப்பான் வாக்கியத்தின் வலப்பக்கத்திலிருந்து தொடங்கி கோவையின் முன்னோக்கிச் செயல்படுவதாகக் கொள்வது அவ்வளவு சரியானதல்ல தெளிவாகக் கூறினால் மொழி இம்மாதிரி செயல்படுவதில்லை இதன் அடிப்படையில் கீழ் - மேல் பகுத்துக் குறிப்பான் உளவியல் அடிப்படையில் சரியானது அல்ல என்று வாதிடலாம்.

வேறுபட்ட ஆக்கங்களின் தொடர்ச்சியைப் பின்பற்றினாலும் ஒரே இலக்கணத்தைப் பின்பற்றுவதனாலட இடது பக்கததிலிருந்து தொடங்கும் பகுத்துக்குறிப்பானும் வலது பக்கத்திலிருந்து தொடங்கும் பகுத்துக்குறிப்பானும் பகுத்துக்குறித்தல் அடிப்படையிலும் பெற்ற விளைவின் அடிப்படையிலும் ஒன்றாகும். ஈந்த வலமிருந்து இடம் என்ற வேறுபடலுக்கும் வலப்பகுத்துக்குறித்தல் அல்லது இடது பகுத்துக்குறித்தல் என்பதற்கும் தொடர்பில்லை.

வலது பகுத்துக்குறித்தலானது கீழ் மேல் பகுத்துக்குறித்தலின் விளைவாகும். இது விதிகளின் வலப்பக்கத்தில் காணப்படும் குறியீடுகளைப் பார்த்து வாக்கிய வடிவுகளைச் சுருக்கும். ஒரு மேல் - கீழ் பகுத்துக்குறிப்பான் இடப்பகுத்துக் குறித்தலை நடைமுறைப்படுத்தும. இது விதிகளதின் இடப்பக்கத்தில் காணப்படும் குறியீடுகளை விரித்து வாக்கிய வடிவுகளை ஆக்கும்.

இதுபோல் ஒரே மொழியியல் செய்திகளின்படி ஒரே வாக்கியங்களுக்கு ஒரே அமைப்புகளைத் தருவதால் ஒரே இலக்கணத்தை நோக்கீடு செய்யும் மேல் - கீழ் பகுப்பானும் கீழ் - மேல் பகுப்பானும் ஒன்றாகும். ஆவைகள் வேறுபட்ட ஆக்கங்களின் தொடாச்சியைப் பின்பற்றுகிறது என்ற உண்மையுடன் எவ்வளவு நினைவகம் தேவை எவ்வளவு கணிப்பு நேரத்தை உட்படுத்துகிறது என்பதாலும் வேறுபடும்.

ஆழம் முதல் மற்றும் அகலம் முதல் பகுத்துக்குறித்தல் (Depth first Versus breath first Parsing)

பின்வரும் ஒரு குழும விரித்தெழுது விதிகளை எடுத்துக் கொள்வோம்.

/ a, S ⟶ AB

/ b, S ⟶ CD

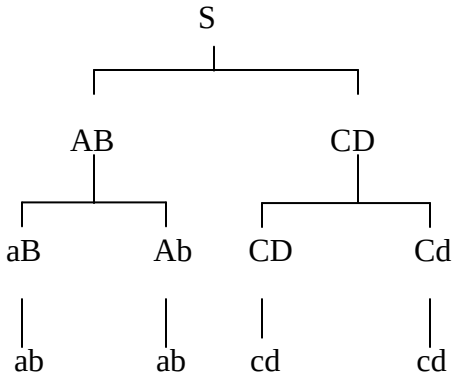
/ c, S ⟶ a

/ d, B ⟶ b

/ e, C ⟶ c

/ f, D ⟶ d

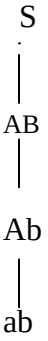
இவற்றோடு தொடர்புடைய பின்வரும் ஆக்கக் கிளையமைப்பையும் எடுத்துக் கொள்வோம்.



இந்த கிளை அமைப்பை இருவிதமாக அணுகலாம் ஒன்று அதன் செங்குத்து நிலையில் கவனம் செலுத்திப் பெறலாம். இரண்டு அதன் இடை நிலையில் கவனம் செலுத்திப் பெறலாம். இந்த இரு வேறுபட்ட நோக்கும் பகுத்துக் குறிக்கும் நடைமுறைத்திறனை பகுப்பாய்வை அளவீட்டை விளைவிக்கும்.

ஆழம் முதல் பகுத்துக்குறித்தல்

நாம் 2 – இல் தரப்பட்டுள்ள கிளை அமைப்பின் செங்குத்து வழியை எடுத்துக் கொள்வோம். வேரை வாக்கியத்துடன் இணைக்கும் ஒரு தனி செங்குத்து விதியைத் தேர்ந்தெடுப்போம். எடுத்துக்காட்டாக கிளைக் கணுவில் அடங்கும் ab என்ற வாக்கியத்திற்கு பின்வருமாறு வழியைத் தேர்ந்தெடுதக்கலாம்.



இந்த வழி கிளை அமைப்பில் வரும் எந்தச் செங்குத்தான வழியைப் போல வாக்கிய வடிவுகளின் ஒரு தொடர்ச்சியைத் தரும்.

ஒவ்வொரு வாக்கிய வடிவும் வேருடனோ அல்லது வேறு தனி விதியால் விளைந்த வாக்கிய வடிவுடனோ ஒரு தனி விதியைப் பயன்படுத்துவதால் விளைவதாகும்.

ஒரு குறிப்பிட்ட ஆக்கத்திற்கு பல விதிகளைப் பயன்படுத்தலாம் என்றாலும் ஒரே ஒரு விரித்தெழுது விதிதான் தக்கவைக்கப்படும் பிற விருப்புகள் கிளையின் பிற வழிகளில் வெளிப்படுத்தப்படும். ஒரு தனிச் செங்குத்தான வழியாகப் பின்பற்ற இயலாது.

ஒரு கிளை அமைப்பில் ஒரு தனி செங்குத்தான வழியிலே உருப்படுத்தம் செய்யத்தக்க வாக்கிய வடிவுகளைப் பின்பற்றும் பகுத்துக் குறிப்பான் ஆழம் முதல் பகுத்துக் குறிப்பான் என்று அழைக்கப்பெறும். மேல் - கீழ் மற்றும் கீழ் - மேல் பகுத்துக் குறிப்பான்கள் இவ்வகையைச் சாரும். கிளைப்படம் 2 ஒரே வாக்கியத்தை ஒன்றுக்கும் மேற்பட்ட வழிகளில் தொடக்கக் குறியீடான 'வா' என்பதுடன் இணைக்கலாம் எனக் காட்டும். ஆழம் முதல் பகுத்துக் குறிப்பான் ஒரு சமயத்தில் ஒரு வழியை ஆய்வதால் தொடக்கத்தில் தேர்ந்தெடுக்கப்பட்ட பாதை தவறாகப் போகலாம்.

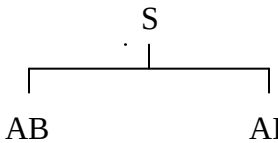
இந்த இடங்களில் பகுத்துக் குறிப்பான் சரிசெய்து தவறிலிருந்து விடுபட வேண்டியது அவசியம் இதன் காரணமாக ஆழம் முதல் பகுப்பான்கள் பின்னழுவல் வசதியுடன் நிறைவேற்றப்பட வேண்டும்.

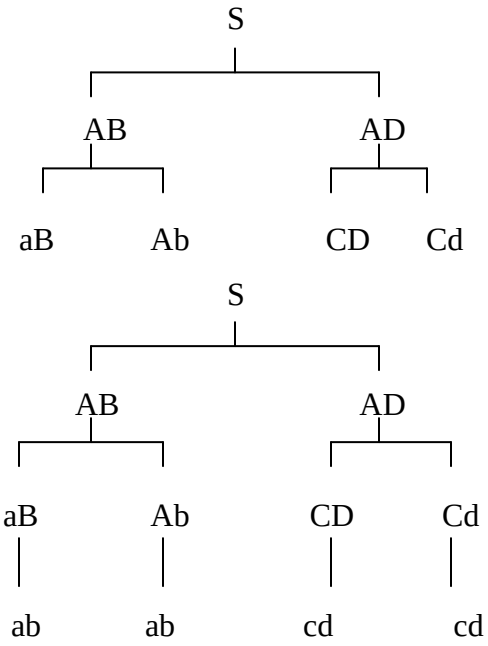
ஆழம் முதல் பகுத்துக் குறிப்பான்

நாம் ஒரு ஆக்கக்கிளை அமைப்பை இடைநிலைப் பரிமாணத்திற்கு முக்கியத்துவம் கொடுத்து கிளை அமைப்பின் ஒரு நிலையில் எல்லாக் கணுக்களையும் எடுத்துக் கொண்டு பார்க்கலாம். எடுத்துக்காட்டாக ஆக்கக்கிளை அமைப்பின் வேர் 'S' என்ற வாக்கிய வடிவைத் தாங்கும் கணுவைக் கொண்டிருக்கும்.

AB மற்றும் AD என்ற வாக்கிய வடிவுகளைக் கொண்ட இரண்டு மகள் கணுக்கள் S என்பதிலிருந்து வேறுபட்ட விதிகளால் விளைந்தனவாகும்.

இம்மாதிரியான நிகழ்வில் பகுத்துக் குறிபர்பான் ஒரே நேரத்தில் வேறுபட்ட ஆக்கங்களை உருவாக்கும். அடுத்த அடியில் இரண்டின் விளைவுகளின் மேல் இயலும் எல்லா விதிகளும் செயல்படுத்தப்படும். இவ்வகையிலான பகுத்துக் குறிப்பான் அகலம் முதல் பகுத்துக் குறிப்பான் எனப்படும்.





வீதி முதல் பகுத்துக் குறிப்பான்கள் உருவாக்கப்பட்ட எல்லா வாக்கிய வடிவிலும் இயலும் எல்லா விதிகளை செயல்படுத்தும் அவை ஆக்கக் கிளை அமைப்பின் இடைநிலைப் பரிமாணத்தை ஆயும் ஒரே நேரத்தில் வரும் எல்லா தேர்வுகளையும் தீர்த்து விட்டு தோல்வி அல்லது வெற்றி என்ற முடிவுக்கு எடுத்துச் செல்லும் இவ் வகையிலான செயல்முறை தோல்வியின் போது பின்னழுவலை அதிக பட்சமாகும். தவறான தேர்வால் விளையும் ஆக்கத் தொடர்ச்சி இல்லாது போனாலும் ஒரே சமயத்தில் உருவான எல்லா வெற்றிப்பிறழ்ச்சிகளும் எஞ்சும்.

வகைப்பாடு (Classification)

	ஆழம் முதல்	**அகலம் முதல்**
மேல் - கீழ்	மேல் _ கீழ் ஆழம் - முதல்	மேல் _ கீழ் அகலம் _ முதல்
கீழ் - மேல்	கீழ் - மேல் ஆழம் - முதல்	கீழ் - மேல் அகலம் - முதல்

நிலைமாற்ற வலைப்பின்னல்களால் பகுத்துக்குறித்தல் (Parsing with Transsition Network)

நிலைமாற்றப் பின்னல் இலக்கணங்கள் கணினி மொழியியலில் மிகவும் வெற்றிகரமாக நடைமுறைப்படுத்தப்படுகின்றது. இதன் அடிப்படைக் கொள்கையில் அமைந்த பல வித்தியாசமான நிலைமாற்றபின்னல்கள் இருக்கின்றன. நிலைபேறு பெற்ற ஒன்று விரிவாக்கப்பட்ட

நிலைமாற்றப் பின்னலாகும். (Augmented transition Network) இது உட்ஸ் (Woods, 1970) என்பவரின் மாதிரியாகும். இது ATN என சுருக்கி அழைக்கப்படும்.

இலக்கணங்களும் தானியங்கிகளும்:

ஒரு குழும விதிகள் ஒரு மொழியின் வாக்கியங்களின் விளக்கமாக எடுத்துக்கொள்வதோடு அந்த மொழியின் கோர்வைகளை ஏற்கும் அல்லது உருவாக்கும். அருவப்பொறியின் (Abstract Machine) விளக்கமாகவும் எடுத்துக்கொள்ளலாம். இம்மாதிரியான பொறிகள் தானியங்கி என்று அழைக்கப்படும். இவற்றின் சாத்தியமான நடத்தை இவைகள் பொருத்தப்படுகின்ற இலக்கணத்தின் விதிகள் உட்கலவை நிலையில் நிர்ணயிக்கப்படுகின்றது.

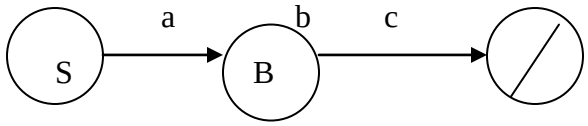
இம்மாதிரியான எல்லா பொறிகளையும் ஒரு குழுமநிலைகளாலும் (Set of States) ஒரு நேரத்தில் ஒரு குறியீட்டை எடுக்கும் உள்ளீட்டுத்திட்டத்தாலும் (Input device) உள்ளீட்டைப் பரிசோதித்து படித்து பொறியை ஒரு நிலையிலிருந்து அடுத்த நிலைக்கு மாறச் செய்யும். ஒரு கட்டுப்பாட்டுத் தொகுதியாலும் (Control Unit) பண்பாக்கம் செய்யலாம். குறியீடுகளை சேகரிக்கவும் பரிசோதிக்கவும் இதைக் கட்டுப்பாட்டுத் தொகுதியில் அடைய இயலும். பொறியின் கணிப்புத்திறன் அதன் நினைவகத்தின் கூட்டுநிலையாலும் நினைவகத்தில் நிறைவேற்றக்கூடிய செயன்மைகளாலும் நிர்ணயிக்கப்படும்.

எளிமையான தானியங்கி முற்றுநிலைப் பொறிகள் (Finite State machins) அல்லது முற்றுநிலைத் தானியங்கிகள் (Finite state automata) எனப்படும். அவைகள் ஏற்கும் மொழியை சீரான மொழிகள் என்பர். சீரானமொழிகளுக்கு சூழல் வரையறை இல்லாத மொழிகளின் (மத்திய, உட்படை/ உட்படுத்தம், விர்த்த) எல்லா பண்புகளும் இருக்கும். இதனால் இதை முற்று எண்ணிக்கை நிலைகள் (Finite Number of States) உள்ள தானியங்கியால் தெரிந்துகொள்ள இயலும் முற்றுநிலைப் பொறிகளுக்கு அக நினைவகம் (Internal Memory) இல்லை. இதன் காரணமாய் அதன் நடத்தை அப்போது படிக்கப்படுகிற குறியீட்டாலும் அப்போதைய பொறி நிலையிலும் நிர்ணயிக்கப்படும்.

இம்மாதிரியான பொறியின் நடத்தையை எடுத்துக்காட்ட பின்வரும் எளிய இலக்கணத்தை எடுத்துக்கொள்வோம்.

$$S \longrightarrow aB$$
$$B \longrightarrow bB$$
$$B \longrightarrow c$$

இது ab * c என்ற சீரான மொழியை உருவாக்கும். இதற்கு சமமான தானியங்கியை மரபு அடிப்படையில் ஒரு நிலைப்படம் (State diagram) முலம் உருப்படுத்தம் செய்யலாம்.



மேலே உள்ள படத்தில் கணுக்கள் நிலைகளைக் குறிப்பிடும். விற்கள் நிலைமாற்றங்களை குறிப்பிடும். விற்களில் உள்ள புலக்குறிப்புகள் (Labels) ஒரு நிலைமாற்றம் சம்பவிப்பதற்கான கட்டுப்பாடுகளைத் தரும். குறுக்குகோடு அடையாளப்படுத்தப்பட்ட கணுவுக்கு ஒரு தனிப்பட்ட நிலை உண்டு. அது இறுதிநிலை எனப்படும். விரித்தொழுது விதிகளிலிருந்து நிலை வரைபடத்திற்கு பெயர்ப்பு செய்வது இந்த தானியங்கிக்கு நேரிடையானது. இறிதியுறா குறியீடுகள் (Non – Lerminal Symbols) நிலைகளுக்குச் (States) செல்லும். இறுதியுற்ற குறியீடுகள் (Terminal Symbols) நிலை மாற்றத்திற்குச் (Transtion) செல்லும்.

இந்த தானியங்கியின் நடத்தையைப் பின்வருமாறு விளக்கலாம். கணித்தல் (Computation) தொடக்கநிலை (Initial Satate) என்று குறிப்பிடப்பட்ட நிலையில் தொடங்குகிறது. (இங்கு S தொடக்கநிலையாகும்). உள்ளீடுகளில் உள்ள முதலாவது குறியீடு அதோடு பொருந்துகிற நிலைமாற்றம் இருக்கிறதா என்று பரிசோதனை செய்யும். அவ்வாறு பொருந்தினால் பொறி நிலை மாற்றத்தைத் தொடர்ந்து அடுத்த நிலைக்குச் செல்லும் மற்றும் உள்ளீட்டுக் குறியை ஏற்றுக்கொள்ளும். அதாவது அது தனது கவனத்தைத் தொடர்ச்சியிலுள்ள அழுத்தக் குறியீட்டிற்கு திரும்பும். இந்த செயற்பாங்கு பொறி நிறுத்துவது வரை தொடர்ந்து நடக்கும். அதாவது கூடுதல் நிலைமாற்றங்கள் இல்லாதது வரை நடக்கும். பொறி இறுதிநிலையில் நிறுத்தும் போது நம்கோர்வையை ஏற்க்கப்பட்டது எனலாம். இதே தானியங்கியை உள்ளீட்டிலிருந்து வெளியீட்டு நிலைமாற்றச் செயலின் பண்பை மாற்றி கோர்வைகளை உருவாக்கச் செய்யலாம். நாம் இறுதி உள்ளே முதல் வெளியே (Last – in – First – Out) பட்டியலுக்குக் கட்டுப்படுத்தப்பட்ட நினைவகத்தையும் (FSA Memory) அதை திறமையாகக் கையாள (Manipulate) கட்டுப்பாடு மரபுகளையும் (Control Convention) சேர்த்தால் தானியங்கியை கீழே தள்ளு தானியங்கியாக (Push Down Autimata) மாற்றலாம். அது கூடுதலான நினைவகத்தின் நன்மை காரணமாக மத்தியில் உட்படையை (Centre embedding) கையாளலாம். கீழே தள்ளு தானியங்கிகள் சூழல் வரையறையற்ற மொழிகளின் (Context Free Language) வகுப்புகளை ஏற்க இயலும்.

மறுதரவு நிலைமாற்ற வலைப்பின்னல்கள் (Recursive Transition Net work (RTNS))

பெரிதாக்கப்பட்ட நிலைமாற்றப் பின்னல் (Augmended transition Network) என்பது நிலைமாற்ற மறுதரவு வலைப்பின்னல் என்று அழைக்கப்படுகின்ற அபூர்வமான சூழல் வரையறையற்ற தெரிந்துகொள்ளும் திட்டத்தின் (Ingenious CF recognition device) விரிவாக்கமாகும். பெரிதாக்கப்பட்ட நிலைமாற்றப் பின்னல்களை இலக்கணங்களின் விரித்தொழுது விதி உருப்படுத்தத்திலிருந்து விரும்பத்தக்க மாற்றமாக அமைகின்றது.

மறுதரவு நிலைமாற்ற வவலைப்பின்னல்களின் அடிபபடைக்காரணம்:

பின்வரும் எளிதான இலக்கணத்தை எடுத்துக்கொள்வோம்.

விதி 1. அ: வா --→ பெ.தொ + வே.தொ * + வி + (து.வி)

விதி 1. ஆ:வா --→ பெ.தொ + பெ.தொ + (து.வி)

விதி 2. பெ.தொ --→ (அ.அ) (அளவடை) பெ.அ * பெ*

விதி 3 : வே.தொ --→ பெ. தொ + வே.உ

இவை சூழல் இல்லாத மொழியை {து.வி, வி, அ.அ, அளவடை, பெ.அ, வே.தொ} என்ற எழுத்துக்கூட்டலால் விளக்குகின்றது. இந்த இலக்கணத்திற்கு குறைந்த ஆக்க சாத்தியம் இருக்கின்றது. பிறை அடைப்புக்குறிக்குள் விருப்பத்தை வெளிப்படுத்தும் நட்சத்திர குறி உள்ள குறியீடு பூஜியத்தையோ மீண்டும் மீண்டும் வருவதையே குறிக்கும். இந்த விதிகளின் வலது பக்க சீரான மொழிகளை விவரிப்பதாக எடுத்துக்கொள்ளலாம்.

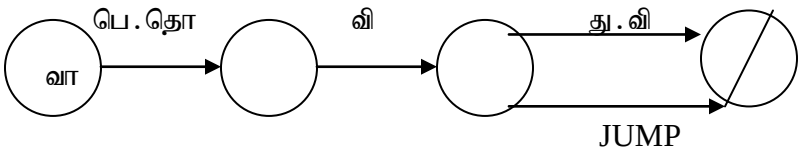
{பெ.தொ, வே.தொ, - வி, து.வி}

{(அ.அ), (அளவடை), பெ.அ * பெ}

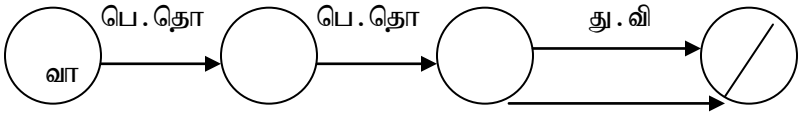
{பெ.தொ- வே.உ}

விதி 1 – இன் இரு பகுதிகளுடன் தொடர்புடைய தானியங்கியின் நிலை வரைப்படங்கள் பின்வருமாறு அமையும்.

அ. வே.தொ



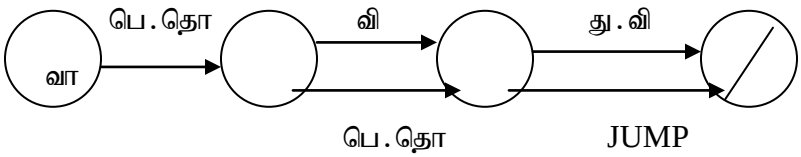
ஆ.



JUMP

அ –வையும் ஆ – வையும் இணைத்து பின்வரும் 1. என்ற வரைபடம் கிடைக்கும். ஆக மேற்கண்ட நான்கு விதிகளுக்கு இணையாக பினவரும முன்று நிலைமாற்ற வரைபடங்கள் கிடைக்கும்.

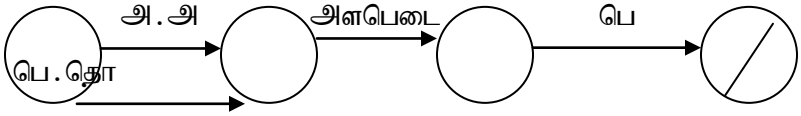
1.



பெ.தொ JUMP

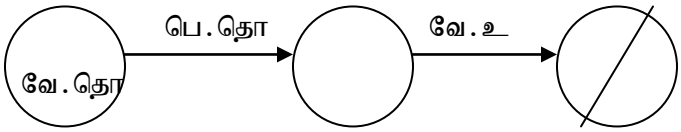
பெ.அ

2.



JUMP

3.



Jumb என்பது எடிஎன் வடிவமைப்பில் உள்ள (Formalism) எந்த உள்ளீட்டையும் ஏற்காத கட்டற்ற நிலைமாற்றத்தை (Unconditional transition) குறிப்பதற்கான இயல்பான மரபாகும்.

எளிய FSA-க்கு வகையான மறுதரவு செயல்பாட்டைக் கையாள இயலாது, ஏனென்றால் அது எங்கே இருந்து வந்தது எவ்வாறு உட்பட்டை கணித்தலுக்குப் பின் எங்குபோவது என்பதை நினைவு கூர்வதற்கு வழியில்லை. FSA –ஐ கீழ்தள்ளு அடுக்கை (Push down Stack) சேர்த்து கீழ் தள்ளு தானியங்கியாக (Push Down Automata –PTA) மாற்றினால் இது சாத்தியமானதாகும். அப்பொழுது இத்திட்டத்தில் கணித்தலின் சேமித்தல் மற்றும் மீட்டல் துணைச் செயன்மைகளைக் கையாளுவதற்குத் தேவையான முலதனம் அமையும். கட்டுப்பாடு தற்காலிகமாக நிறுத்தப்பட்டு மீண்டும் தொடர்வதால் இது இறுதி உள்ளே – முதல் -வெளியே (Last in first out) என்ற அடிப்படையில் நிகழும்.

குறியீடுகளை (Symbols) புலக்குறிப்புகளாகச் (Labels) செயல் அடிப்படையில் விளக்குதல் நிலைபெறு ATN கருத்துச்சாயளில் பிரதிபலிக்கின்றது. உட்படுத்தும் செயற்பாங்கிற்கு கட்டளை தரப்பட்டால் அப்போதைய பொறிநிலை அடுக்கில் தள்ளப்படும். இவ்வாறு கட்டளைத் திரும்பும்போது கணித்தல் அதே இடத்திலிருந்து திரும்பத் தொடங்கப் பெறலாம்.

ATN வடிவவாதம் (Formolism) ஒரு நிலைக்கு கட்டுப்பாட்டின் பெயர்ச்சி எடுத்துக்காட்டாக PUSHNP (பெயர்த்தொடரைத் தள்ளு) என்ற அடையாளப்படுத்தப்பட்ட நிலைமாற்றத்தால் எழுப்பப்படும். இதற்கு பொருத்தமான இறுதி நிலை மரபு POP என்று அடையாளப்படுத்தப்பட்ட பொய்யான வில்லை (ArC) செயல்பாடு நிற்கும் போது முலப் பொறிநிலை அடுக்கிலிருந்து திரும்பவும் தொடக்கநிலை நிறுத்தப்படும்.

நாம் இன்னும் பிற நிலைமாற்றங்களின் (து.வி, வி, அ.அ போன்றவற்றின்) மதிப்புநிலையை நிர்ணயிக்க வேண்டும். நமது விருப்பம் இலக்கண வகைப்பாட்டின் பெயர்களைப் பகுத்துக் குறித்தலாக இருந்தால் ஒழிய அவை உண்மையான உள்ளீட்டு குறிகளாக இருப்பது அரிது.

நாம் நிலைமாற்ற வலைப்பின்னல் (Transition Network) இலக்கணங்கள் எவ்வாறு அகராதிகளைக் கையாளுகின்றன என்பதை குறிப்பிட வேண்டும்.

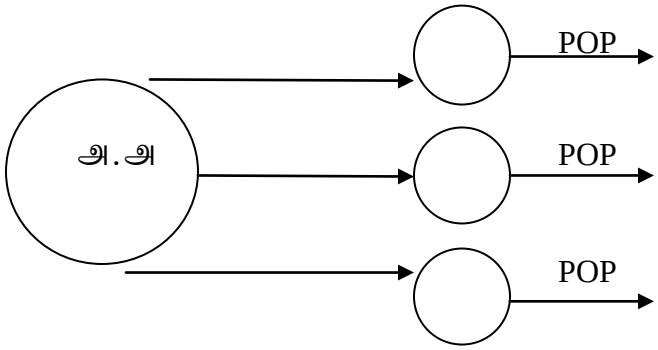
ஒரு சாத்தியம் அகராதியை விரித்தொழுது விதிகளின் (Rewriting rules) தொகுதியாக வெளிப்படுத்த இயலும்.

வி. 4. அ அ.அ ⟶ 'ஒரு'

வி. 4. ஆ அ.அ ⟶ 'அந்த'

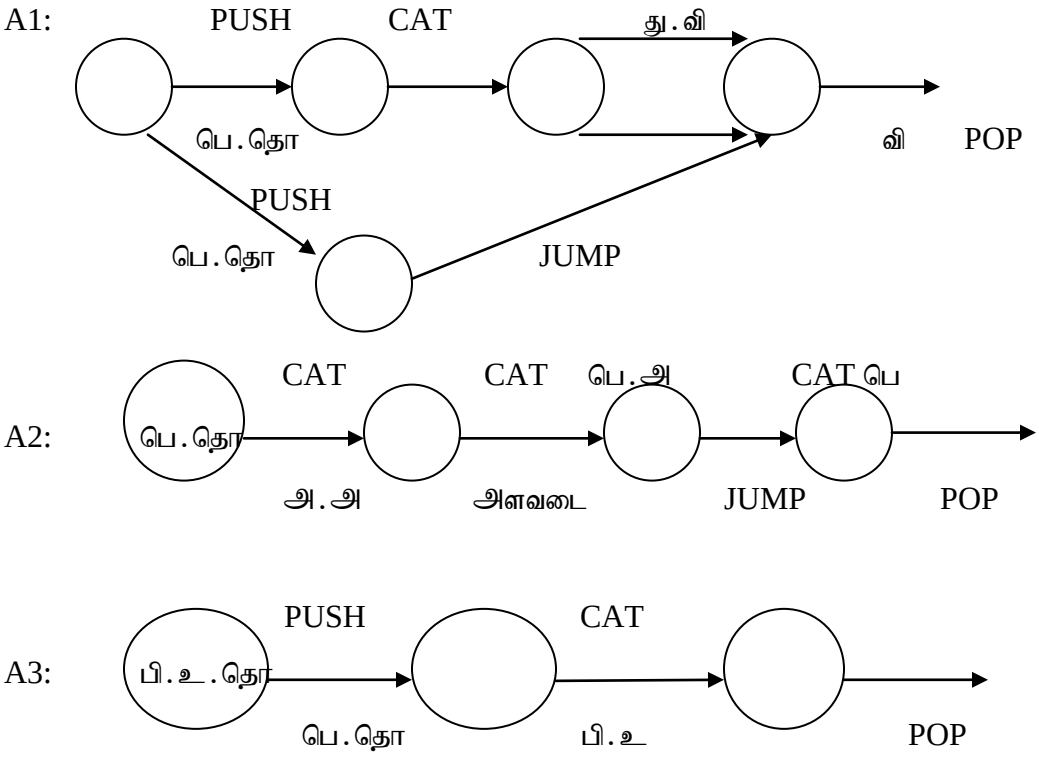
வி. 4. இ அ.அ ⟶ 'கொஞ்சம்'

நாம் சில நுணுக்கங்களைப் பயன்படுத்தி விரித்தெழுது விதிகளை வடிவில் பின்வருமாறு மாற்றிக் கூறலாம்.



மாறாக ATN – இல் எப்போதும் நடப்பதுபோல் நாம் நிலை மாற்றத்தைத் தூண்டும் ஒரு குழும செயல்பாடுகளை விரிக்கலாம். அது முன் விவரிக்கப்பட்ட அகராதியைப் பார்க்கும். அகராதியின் வடிவ அமைப்பு நிலைமாற்ற இலக்கணம் உள்ளடக்கம் கூடுதல் மொழிச் செயல்பாடாக மாறுகிறது. பரவலாகப் பயன்படுத்தப்படும் அகராதியைப் பார்க்கும் செயல்பாடுகள் (பொதுவாக CAT என்று அழைக்கப்படும்) முன் விளக்கம் பெற்ற

வகைப்பாட்டை (Predifined category) அப்போதைய உள்ளீட்டுச் சொல்லுக்கான (Current input word) அகராதி சொல்லுடன் (பொருத்தி) பரிசோதிக்கின்றது. நாம் வலைப்பின்னல் இலக்கணத்தை (Network grammar) (A1-A3) சொல்வகைப்பாட்டுச் சோதனைகளுக்கிடையில் (CAT) மாற்றங்களை வேறுபடுத்துகின்ற மற்றும் (Push, தள்ளு) என்ற துணைச் செயல்பாடுகளை அழைக்கின்ற மற்றும் பொய் POP விற்களால் அடையாளப்படுத்தப்பட்ட இறுதிநிலைகளுடன் கூடிய RTN க்கு நிலைபெறு கருத்துச்சாயலுடன் கூடுதல் நெருக்கமாகப் பொருத்தமுறுகின்ற ஒரு வடிவில் உருப்படுத்தம் செய்யும் நிலையில் நாம் இருக்கிறோம்.

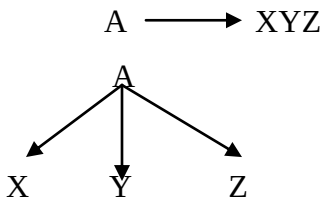


மேற் சொன்னது சூழல் வரையறை இல்லாத விரித்தொழுது விதிகளுக்கு சமமான அருவநிலையாகும். அதாவது சூழல் வரையறை இல்லாத இலக்கணத்தை அதற்குச் சமமான RTN –ஆக மாற்றலாம். இருப்பினும் விரித்தொழுது விதி இல்லாத முக்கியமான RTN பண்புக்கூறு 'செயற்பாங்கு' (Process) என்ற கருத்து சாயலாகும். நாம் ஒரு மொழியின் தெரிந்து கொள்ளும் செயல்முறையை (Recognition Process) உருவாக்க ஒரு இலக்கணத்தை பயன்படுத்த விரும்பினால் விரித்தொழுது விதி என்ன செய்ய வேண்டும் என்று கூறுகின்றது. RTN – இல் அடங்கும் தானியங்கியும் எவ்வாறு செய்ய வேண்டும் என்று கூறுகின்றது. எ.கா.

விரித்தெழுது விதி அடிப்படையிலான சூழல் வரையறையில்லாத 'தெரிந்துகொள்ளலுக்கு' உருவாக்கு ஒழுங்குமுறையும் (Production System) அதே மொழிக்கு 'தெரிந்து கொள்வானாக' நடைமுறைப்படுத்தப்பட்ட RTN ஐயும் ஒப்பிட்டுப்பார்க்கவும் RTN வடிவவாதம் (RTN Formalism) கணினி உள்ளீடுகளின் சில வரிசையான உருபடுத்தம செய்யும் போது மரபு செயல்முறை மொழிகளின் (Conventional Procedured Programming Languages) பண்புகளை வெளிப்படுத்துகின்றது. கட்டுண்டவைகளை மற்றும் கட்டுப்படாதவைகள் இவற்றின் go to நிலை மாற்றங்களும் மொழிபு அடையாள குறிகள் (நிலையின்பெயர்கள்) செயற்பாங்கு கட்டளைகள் (Procedurce cells) (Push: POP) இந்தக் கருத்து சாயலில் முக்கியமான வெளிப்படையாக இல்லாத செயற்பாங்கு 'விளங்கிக்கொள்வான்' (Interpreter) அதிலிருந்து ஒன்றுக்கும் மேற்பட்ட நிலைமாற்றம் வரும் நிலையை அடைகையில் அதன் நடத்தை ஆகும்.

நிலைபெறு ATN கள் (Standart ATNS)

RTN பகுத்து குறித்தல் மிக பயனுள்ள பொறி அல்ல. இது சூழல் வரையறையில்லாத மொழிகளை மட்டும் தான் தெரிந்துகொள்ளும். அமைப்பாக்கம் செய்யப்பட்ட வெளியீட்டை தருவதற்கான வழிமுறை இல்லை. உறுபப்மைப்பு இலக்கணங்கள் என்று அறியப்படுகின்ற விரித்தொழுது விதி ஒழுங்குமுறைகளின் ஓரளவுக்குப் பெரிய துணைவகுப்புக்கு விரித்தெழுது அடையாளக் குறிகளை படிநிலை ஆதிக்க (Hierachical dominece) இயக்கியாக (Operator) மீட்டுருவாக்கம் செய்து ஒரு புரிந்து கொள்ளுதல் செயல்பாட்டிலிருந்து வெளியீடாக ஒரு பகுத்துக்குறித்தல் கிளை அமைப்பை உருவாக்க இயலும்.



RTN – ஐ சூழல் கட்டுண்ட பகுத்துக்குறித்தலாகப் பயன்படுத்த கடுமையாக திட்டமிடல் வேண்டும். வடிவவாதம் ஒரே நேரத்தில் ஒன்றுக்கும் மேற்பட்ட குறியீடுகளை (Symbols) பரிசோதிக்க வழி தராத காரணத்தால் கட்டுண்ட தொடரமைப்பு விதிகளின் (Context sensitive Phrase Structure rules) பின்வரும் நிலைபேறு வடிவங்களில் ஒன்றை நிலைமாற்ற வலை அமைப்பு (Transition Network) உருப்படுத்தமாக பதிலீடு செய்ய நேரடியான வழி இல்லை.

$$XAY \longrightarrow X\,BC\,Y$$

அல்லது

$$A \longrightarrow BC/X\text{-}Y$$

முற்றுநிலை தெரிந்து கொள்வானின் (Finite State recognizer) கரடுமுரடான கட்டமைப்பிலிருந்தான (Crude arctitecture)

மாதிரி வடிவில் பரம்பரையாகப் பெற்ற விளக்கங்களின் பரம்பரை மேல் - கீழ் (Top – Down) இடமிருந்து வலம் (left – right) தொடர்ச்சி நுணுக்கங்களின் இவ்வகையிலான விதிகளை நினைத்துப்பார்ப்பதைக் கூட நிலைமாற்ற வலை அமைப்பு சட்டகத்தில் விகாரமாகச் செய்கிறது. சூழல் கட்டுண்டமையின் மேல் - கீழ் இடமிருந்து வலம் நடைமுறை எது பின்னால் வருகிறது என்று அறியும் சந்தர்ப்பம் கிடைப்பது வரை நிச்சயங்களைப் பின்தள்ளிவைக்கச் சாத்தியப்பட வேண்டும். நிலைமாற்ற வலை அமைப்பின் இயந்திரத்தன்மையான பரிமாணம் அடிப்படையில் பின்னால் வரும் சில நிலை மாற்றங்களுக்குக் (transition) கட்டுப்பாடுகளைக் கூடுதலாகத் தர முந்திய நிலைமாற்றங்களின் விவரணைகளைச் சேமித்துவைக்கவும் மீட்கவும் கூடுதலான நினைவகம் தேவை. ATN – இல் வலை அமைப்பின் ஒவ்வொரு நிலையிலும் ஒரு குழுமப்பதிவுகளை அவற்றை பயன்படுத்தும் வண்ணம் சில முன்விளக்கப்பட்ட செயல்பாடுகளுடன் சேர்க்கின்றோம். நிலைபேறு ATN – இல் பதிவுகளில் தரப்பட்டுள்ள மிக முக்கியமான செயன்மைகள் (OPerations) SETR என்ற ஒப்படைக்கும் செயல்பாடும் ஒரு பதிவுக்கு தற்போதைய மதிப்பைத் தரும் GETR என்ற செயல்பாடும் ஆகும்.

பதிவுகளையும் (Registers) அது தொடர்பான செயன்மைகளையும் சேர்த்து RTN –ஐ சூழல் கட்டுண்ட மொழிகளைத் தெரிந்துகொள்ளச் செய்யலாம். இருப்பினும் இலக்கணம் வழியாக பயணத்தின் தடத்திலிருந்து தானாகவே பகுத்துக் குறித்த கிளை அமைப்பை (Parse frea) உருவாக்கும் முக்கியமான பண்பை இழந்துவிடலாம். இருப்பினும் ATN Literals மற்றும் பதிவுப்பொருளடக்கத்திலிருந்து எந்த பட்டியல் அமைப்பை உருவாக்குவதை அனுமதிக்கும் அமைப்பு - உருவாக்குச் செயல்பாடான (Structure bullding fundion) BUILD 2 -ஐத் தருகிறது. மேலும் இந்த இடுகுறித்தன்மையான பட்டியல் கிளை அமைப்புகளாகப் புரிந்துகொள்ளப்படுகின்றன. கிளை அமைப்பின் எந்த வகையையும் காட்டவும் சேகரிக்கவும் BUILD 2 – இன் பயன்பாட்டால் சாத்தியமாகும். இதை வெளியீடு செய்யவும் வசதி வேண்டும். ATN கட்டளைச் செயல்பாட்டிற்கு (Celling Process) ஒரு மதிப்பைத் திருப்பித்

தர POP –ஐ அனுமதித்து வெளியீட்டைத் தருகிறது. பதிவுகள் வெளியீடு மற்றும் பதிவுகளின் மேல் சோதனைகளால் நிலைமாற்றங்களைக் கட்டுப்படுத்தும் ஆற்றல் இவற்றில் RTN ATN ஆக மாறுகிறது. சொற்களின் தொடர்ச்சியை கிளை அமைப்பாக மாற்றுகிறது.

செயல்முறை கட்டளைகள்

1. (.CAT.) : வகைப்பாட்டுப் பரிசோதனையை செயல்படுத்தும் காரிய கிரமம்.
2. (.PUSH.) : தள்ளுப் பரிசோதனையை செயல்படுத்தும் காரியக் கிரமம்.
3. (.TST.) : நிலைமாற்ற அட்டவணைப் பரிசோதனையைச் செயல்படுத்தும் காரிய கிரமம்
4. (.POP.) இறுதியைத் தீர்மானித்து BULD ஐக் கூப்பிடும் காரிய கிரமம்.
5. (.SETR.) பதிவின் மதிப்பை நிறுவும் காரியக் கிரமம்.
6. (.SENDR.) கீழ்நிலைக்கு பதிவு மதிப்பை அனுப்பும் காரியக் கிரமம்.
7. (.CIFTER.) மேல்நிலைக்கு பதிவு மதிப்பை திருப்பும் காரியக் கிரமம்.
8. (.TO.) உள்ளீட்டுக் காட்டியை முன்னேற்றி நடைமுறைபடுத்த வேண்டிய அடுத்த கணுவை நிர்ணயிக்கும் காரியகிரமம்.
9. (.JUMP.) நடைமுறைப்படுத்த வேண்டிய அடுத்த கணுவை நிர்ணயிக்கும் காரியக் கிரமம்.
10. (.GETR.) பதிவின் மதிப்பை பெறும் காரிய கிரமம்.
11. (ooo) நட்சத்திரப் பதிவின் மதிப்பை பெறும் காரியகிரமம்.
12. (.GETF.) அகராதியிலிருந்து பண்புக்கூறை மீட்டும் காரிய கிரமம்.
13. (.APPEND.) ஒரு தரவு சாதனத்தைப் பதிவில் சேர்க்கும் காரிய கிரமம்
14. (.BUILD.) புதிய அமைப்பை உருவாக்க வேண்டி பல பதிவுகளில் இருக்கும் மதிப்புகளை இணைக்கும் காரிய கிரமம்.

பாணினி இலக்கண வடிவவாதம்

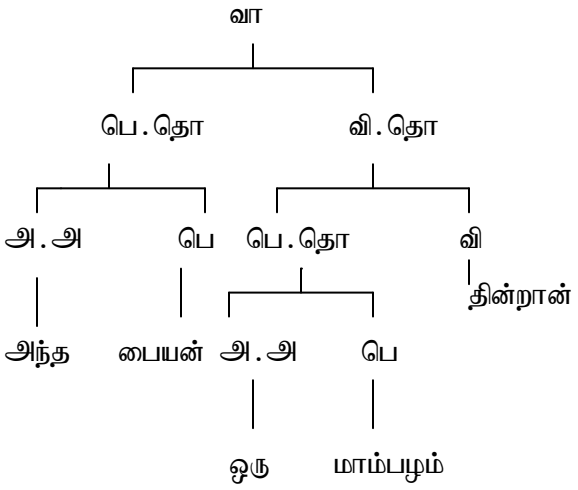
இலக்கண வடிவவாதம்

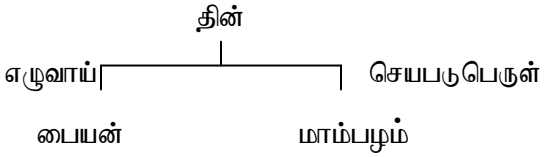
இந்திய மொழிகளின் பகுப்பாய்வுக்கும், பகுத்துக் குறித்தலுக்கும் பாணினியின் வடிவவாதம் மிகப் பரவலாகப் பயன்படுத்தப்படுகின்றது. இவ்வடிவ வாதத்தால் இந்திய மொழிகளுக்குப் பிற வடிவவாதங்களை விட பாணினியின் வடிவவாதம் பொருத்தமானது என வாதிடுகின்றனர்.

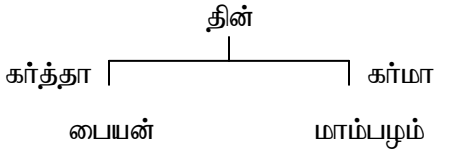
கூற்றுக்கும் கேட்பவர் எடுத்துக் கொள்கிற பொருண்மைக்கும் இடையே உள்ள உறவை நிலைநிறுத்தும் விதிகளின் ஒழுங்குமுறை (இயற்கை மொழியின்) இலக்கணமாகும். எடுத்துக்காட்டாக பின்வரும் வாக்கியத்தை எடுத்துக்கொள்வோம்.

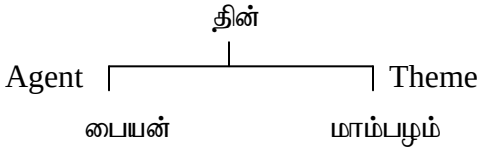
அந்தப் பையன் ஒரு மாம்பழம் தின்றான்

வழி 1









இலக்கண வடிவவாதம் 2

விரும்புகிற பொருளைத் தருகிற சொற்குழுக்களுக்கு இடையே உறவைப் பிரித்தெடுக்கும் நுணுக்கத்தை தருவது இந்த இலக்கண வடிவவாதத்தின் வேலையாகும்.

பாணினி இலக்கண வடிவவாதம்

இவ்வடிவ வாதம் பின்வரும் முக்கியமான கேள்வியை எழுப்பபுகின்றது

கூற்றிலிருந்து பொருண்மையைக் கேட்பவர் எவ்வாறு பெறுகிறார்?

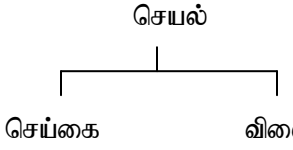
அல்லது

எவ்வாறு மொழி பொருண்மையைக் குறியாக்கம் செய்கின்றது?

- எந்த வடிவில்
- எங்கே
- எவ்வளவு

பாணினி இலக்கணத்தின் பொருண்மை மாதிரி

பாணினி இலக்கணத்தின் பொருண்மை மாதிரி செயலை செய்கையாகவும் விளைவாகவும் பகுக்கின்றது.



பாணினியின் பொருண்மை மாதிரியில் முக்கியமான கருத்துச்சாயல் காரகம் ஆகும். காரகம் செயலுக்கும் அதில் பங்கு பெறுவோர்க்கும் இடையே உள்ள உறவை வெளிப்படுத்தும் இம்மாதிரியில் செய்கையானது பல துணைச் செய்கைகளாகப் பிரிக்கப்படும்.

செய்கை

துணைச் செய்கை 1

துணைச் செய்கை 2

துணைச் செய்கை 3

துணைச் செய்கை 4

பேசுபவரின் விருப்பம் இம்மாதிரியில் வெளிப்படும் இம்மாதிரி பின்வரும் பண்புக் கூறுகளால் இந்திய மொழிகளுக்குப் பொருத்தமாக அமைகின்றது.

1. சுத்தமான சொல் வருகை முறை
2. வேற்றுமை உருபுகள் காரணமாக இருத்தல்

பெயருடனான வினையின் உறவு வேற்றுமை உருபுகளாகும். வாக்கியத்திலிருந்து எவ்வாறு காரக உறவுகளைப் பிரித்தெடுப்பது? என்ற கேள்விகளுக்கு விடை இரண்டு சிக்கல்களை நேரிடும்.

முதல் சிக்கல்

வேற்றுமை உருபு - காரகம் பல – ஒன்று பொருத்தம் காட்டும்

எடுத்துக்காட்டு

1. இராமன் மாங்காய் தின்கிறான்

2. இராமன் மாங்காயைத் தின்கிறான்

3. இராமனுக்கு மாங்காய் தின்ன வேண்டும்

4. இராமனால் மாங்காய் தின்னப்பட்டது

5. இராமனால்/ இராமனுக்கு மாங்காய் தின்னமுடியவில்லை

மேற்சொன்ன எல்லா வாக்கியங்களிலும் இராமன் - கர்த்தா (எழுவாய்) ஆனால் வேற்றுமைகள் வேறுபட்டவை.

தீர்மானம்

வேற்றுமை உருபுகள் செயல்பாடு (வினையின் காலம், ஆற்றல்) நோக்கு சார்ந்தது. குறிப்பாக வேற்றுை உருபு - காரகப் பொருத்தம் வினையிலிருந்து சுதந்திரமானது மற்றும் வினையின் காலம், ஆற்றல், நோக்கு இவற்றைச் சார்ந்தது.

உருப்படுத்தம்

முன்னர் தரப்பட்ட எடுத்துக்காட்டு அடிப்படையில் தின் என்ற வினைக்கான காரக அட்டவணை பின்வருமாறு அமையும்.

எடுத்துக்காட்டு

அ) தின் என்ற வினைக்கான காரக அட்டவணை

காரகம்	**தேவை**	**வேற்றுமை உருபு**
கர்த்தா	கட்டாயம்	0
கர்மா	கட்டாயம்	0/ஐ
காரணம்	விருப்பம்	ஆல்

(ஆ) மாற்று விதிகள்

வினையின் காலம், ஆற்றல், நோக்கு அதன் அடிப்படையில் மாற்றங்கள் ஏற்படும்

மாற்றங்கள அ - வேண்டும்

கர்த்தா : 0 > உக்கு

அ - பட்டது கர்த்தா : 0 > ஆல்

அ - முடியவில்லை/ முடியும் கர்த்தா : 0 > உக்கு/ ஆல்

இரண்டாவது சிக்கல்

வேற்றுமை உருபுக்கும் காரகத்திற்கும் இடையில் ஒன்று பல பொருத்தம் வரும்.

எடுத்துக்காட்டு

ராமன் சாவியால் பூட்டைத் திறந்தான்

சாவி பூட்டைத் திறந்தது

பூட்டு திறந்தது

வேற்றுமை உருபு + 0

மேற்சொன்ன எடுத்துக்காட்டில் எழுவாயின் காரகம், கர்த்தா, காரணம், கர்மா என்ற மூன்று நிலைகளில் வருகின்றது. எந்தச் செயலின் காரகம் என்பது இச்சிக்கலுக்குத் தீர்வைத் தரும்.

தீர்மானம்

அகப் பொருண்மை மாதிரி பின்வருமாறு பகுத்தாயப்பட வேண்டும்.

பூட்டுத் திறத்தல் (முக்கிய செயல்)

1. சாவியைச் செருகித் திருப்புதல்
2. சாவியை நெம்புகோலில் அழுத்தி அதை நகரச் செய்வது
3. தாழ் பின் வாங்குதல் மற்றும் பூட்டுத் திறத்தல்

ராமன் : தொடக்கச் செயல்பாடு ஒன்றின் காரகம்

சாவி : தொடக்கச் செயல்பாடு இரண்டின் காரகம்

பூட்டு : தொடக்கச் செயல்பாடு மூன்றின் காரகம்

காரகப் பங்களிப்புகளைத் தருவதில் வாக்கிய கட்டுப்பாடுகள் பின்வருமாறு அமையும்

1. ஒரு வினைக் குழுவின் ஒவ்வொரு கட்டாயக் காரகமும் சரியாக ஒரு மூலச்சொல் குழுவுக்குத் தரப்பட வேண்டும்.
2. ஒரு வினைக் குழுவின் ஒவ்வொரு விருப்புக் காரகமும் சரியாக குறைந்தது ஒரு மூலச் சொல் குழுவுக்குத் தரப்பட வேண்டும்.
3. ஒவ்வொரு மூலச் சொல் குழுவும் சில வினைக்குழுவில் சரியாக ஒரு காரகத்திற்குத் தரப்பட வேண்டும்.

பில்மோர் (Fillmore, 1960) காரகக் கோட்பாட்டைப் பின்வரும் நான்கு நிலைகளாகக் குறிப்பிடுகிறார். அவையாவன :

பொருண்மை நிலை (பேசுபவர் மனதில்)

காரக நிலை

வேற்றுமை உருபு நிலை (வேற்றுமை உருபு)

புறநிலை (வாக்கியம்)

எடுத்துக்காட்டு, பொருண்மை நிலை

தின் : செயல் ராமன் R1, மாங்காய்: R2, கை : R3

காரக நிலை

தின் - இநற்தகாலம், ராமன் : K1, மாங்காய் K2, கை : K3

வேற்றுமை நிலை (வேற்றுமை உருபு)

தின் : த் ராமன் : மாங்காய் : கை: ஆல்

புற நிலை (வாக்கியம்)

பாணினியின் பகுத்துக் குறித்தல் (Paninian Pansor)

பாணினியின் பகுத்துக் குறித்தல் அடிப்படையில் ஒரு மொழியின் சொற்குழுமங்களையும் இலக்கணத்தையும் பயன்படுத்தி முக்கியப் பகுத்துக் குறிப்பான் வாக்கியத்தின் பகுத்துக் குறித்தலைச் செய்யும் இதைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.

மொழி L – லிருந்து சொற்குழுக்கள்

மொழி L –ன் இலக்கம்_____ முக்கியப் பகுத்துக் குறிப்பான்

வாக்கியத்தின் பகுத்துக் குறித்தல்

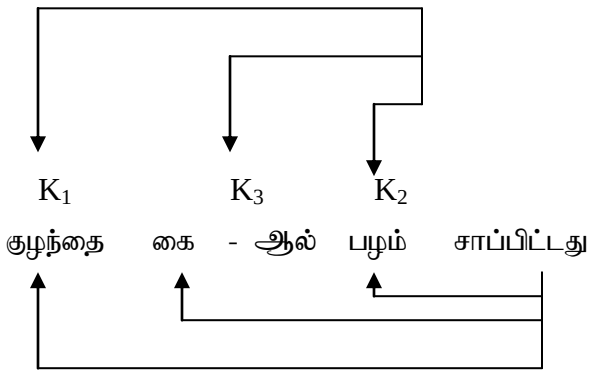
பகுத்துக் குறித்தல் பின்வருவனவற்றை வேண்டும்.

1.(அ) வினை வேர்களுக்கான காரக அட்டவணை

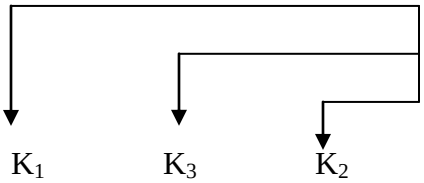
(ஆ) மாற்று விதிகளின் காரக அட்டவணை

2. காரகப் பாத்திரங்கள் தரப்படுவதன் மேலான வாக்கிய கட்டுப்பாடுகள் ஆகியவையாகும்.

காரக அட்டவணையைப் பயன்படுத்தி நாம் பின்வரும் கட்டுப்பாடு வரைபடம் பெறுகின்றோம்.

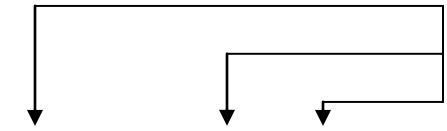


வாக்கிய கட்டுப்பாடுகளைப் பயன்படுத்தி நாம் இரண்டு தீர்மான வரைபடங்களைப் பெறுகின்றோம்.



K_1 K_3 K_2

குழந்தை கை - ஆல் பழம் சாப்பிட்டது



K_2 K_3 K_1

குழந்தை கையால் பழம் சாப்பிட்டது

K_1	கர்த்தா
K_2	கர்மா
K_3	காரணம்

முக்கியப் பகுத்துக் குறிப்பான் (Core Parser)

முக்கியப் பகுத்துக் குறிப்பான் பின்வரும் செயல்பாடுகளைச் செய்யும்.

1) சொற்களுக்கு இடையில் காரக உறவை அடையாளம் காணும் (எழுவாய்க்கும், வினைக்கும் இடையே உள்ள உறவு)

2) சொற்களின் பொருண்மைகளை அடையாளம் காணும்

   1. காரக – வேற்றுமை உருபு பொருத்தம்
   2. காரகத்தின் விருப்பம்
   3. மாற்று விதிகள்

குறிப்பிட்ட இடம் சார்ந்த அதாவது வட்டாரச் சொற்களை குழுமுதல் (Local Word Grouping)

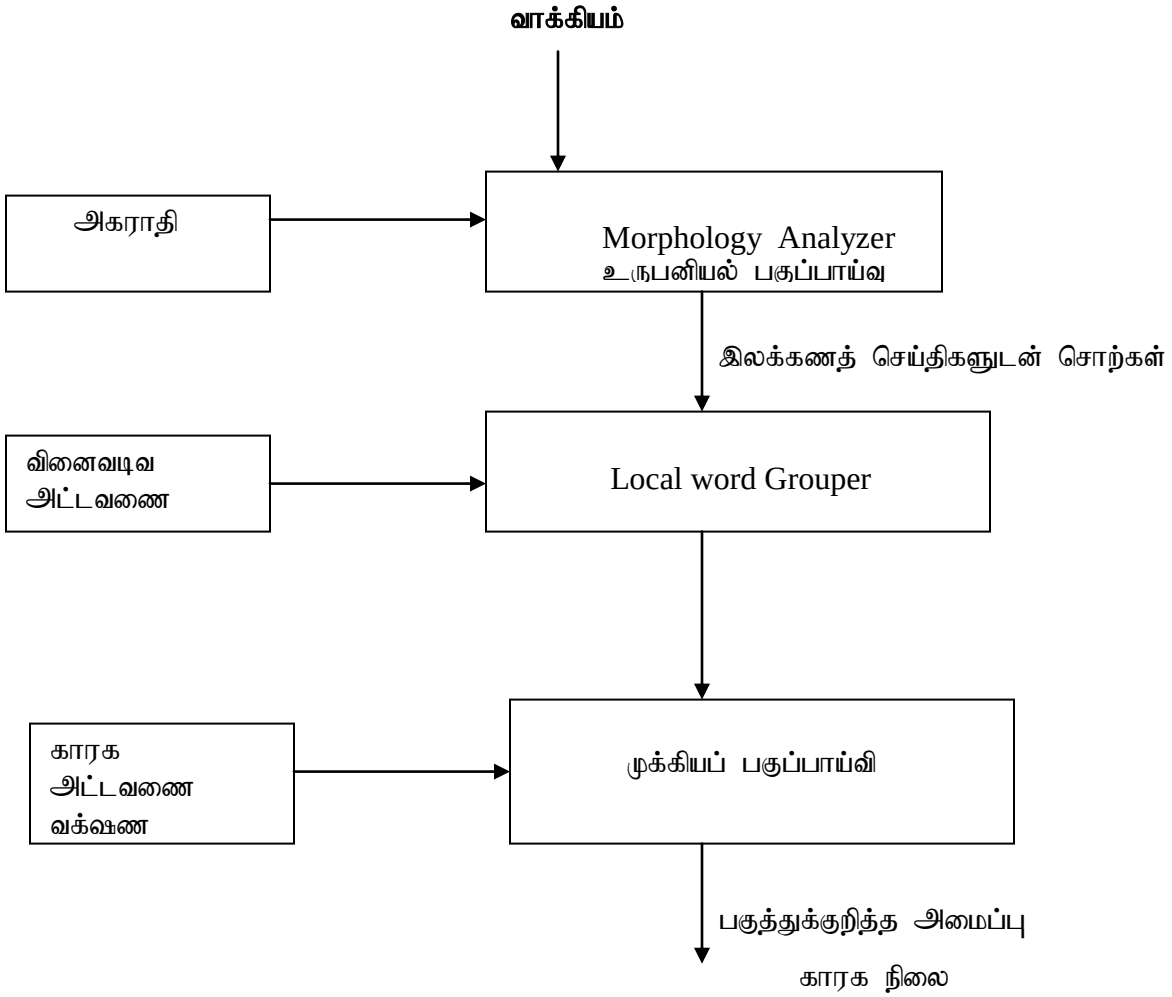
இது பாணினி பகுத்துக் குறிப்பானில் முக்கியமான கருத்துச் சாயலாகும். இதன் படி சொற்கள் பகுத்துக் குறித்தலுக்காக ஒன்றாகக் குழுமப்படும். அனால் இச்செயல்பாடு தொடரமைப்பு இலக்கணங்களில் செய்யப்படும் 'வாக்கியங்களைத் தொடர்களாகப் பிரித்தல்' என்பதிலிருந்து பின்வரும் உண்மைகளின் அடிப்படையில் வேறுபடும்.

- இந்திய மொழிகளில் வினைத்தொடர் இயற்கையாக அமையத் தேவையில்லை.
- கணித்தல் அடிப்படையில் பெயர்த் தொடர்களையும் வினைத் தொடர்களையும் தெரிந்து கொள்ளுதல் அவ்வளவு எளிதானதோ வலுவானதோ அல்ல.
- எனவே பெயர்க்குழுக்களும், வினைக் குழுக்களும் புறச் செய்திகளைக் கொண்டு செயல்படுகின்றது.
- முக்கியமானது வினைக் குழுக்களின் பொருண்மைச் சாயல்களின் தெரிதலாகும்.

பின்வருவன முக்கியமான தகவல்களாகும்.

1. வினை வேர்களின் சாத்தியமான தொடர்ச்சி
2. பால் - எண் - இட உடன்பாடு

இதுவரைச் சொன்ன செய்திகளின் அடிப்படையில் பாணினி பகுத்துக் குறிப்பானைப் பின்வருமாறு உருப்படுத்தம் செய்யலாம்.



பகுத்துக் குறித்தலின் முக்கியத்துவம்

பகுத்துக் குறித்தலில் காரகங்களின் முக்கியத்துவம் பின்வருமாறு அமையும்.

- கர்த்தாவுக்கு கீழ்நோக்கு வரிசையில் பின்வருமாறு முக்கியத்துவம் அமையும்
- மனிதன், மனிதன் அல்லா உயிர்கள், உயிரில்லாதவை
- மூலக் குழுமம் அது விரும்புகிற குழுமத்துடன் சேர்ந்து வரும்
- கர்த்தா கர்மாவுக்கு முன்வரும்.

பொருள் மயக்கத்தைப் போக்க லக்ஷண அட்டவணை வாக்கியத்தில் உள்ள வினையின் பொருண்மையைக் காரண வினைகளின் காரகத்தைப் பயன்படுத்துகின்றது.

எடுத்துக்காட்டு

அவன் அவளிடம் மன்னிப்புக் கோரினான்

அவன் கிணற்றிலிருந்த தண்ணீர் மொண்டான்

அவன் அவளிடம் பணம் கேட்டான்

அவன் அவள் சொல்வதைக் கேட்டான்

வினை லக்ஷண அட்டவணை அகராதிகளைப் பயன்படுத்தி மொழியியலாளர்கள் உருவாக்கப்படும் பெயர் லக்ஷண அட்டவணை ஒரு வாக்கியத்தில் வரும் பெயர்களின் பொருண்மை மயக்கங்களைப் போக்க உதவும்

எடுத்துக்காட்டு

அவன் வயலை உழுகிறான்

அவன் வயலை அறுக்கிறான்

சொல் வகுப்புகளும் சொல் வகைப்பாட்டு அடையாளம் படுத்துகையும் (Word class and Part of Speech Tagging)

மொழி ஆய்வில் சொல் வகைப்பாட்டின் முக்கியத்துவம் என்னவென்றால் அது சொல்லைப்பற்றியும் அவற்றை அடுத்து வரும் சொற்கள் பற்றியும் செய்திகளை ஏராளமாகத் தருகின்றது. இவை முக்கிய வகைப்பாடுகளுக்கும் உண்மையாகும். (வினை *பெயர்) மட்டுமின்றி பல நுட்பமான வேறுபாடுகளுக்கும் உண்மையாகும். எ.கா இந்த அடையாளக் குழுமங்கள் உடமை பற்றும் பெயர்களுக்கும் (என்னுடைய, உன்னுடைய, அவனுடைய, அவளுடைய, அவருடைய, அதனுடைய) முவிடப்பெயர்களுக்கும் (நான், நீ, அவன், அவள், அது) வேறுபாட்டைக் காட்டுகின்றன.

ஒரு சொல் உடமை மாற்றுப்பெயரா அல்லது முவிடப்பெயரா என்பதை அறிந்து கொண்டு எந்தச் சொற்கள் அதன் பக்கத்தில் வர இயலும் என்பதை கூற இயலும். இதை பேச்சு தெரிந்துகொள்ளுதல் (Speech recognition) என்பதற்கான மொழி மாதிரியில்

(Language Mode) பயன்படுத்தலாம். சொல்வகைப்பாடு தகவல் மீட்டலுக்கான (Information reterival) வேர்ச்சொல்லை கண்டுபிடித்தலுக்கு (Stemming) பயன்படுகின்றது.

ஏனென்றால் சொல் வகைப்பாட்டை (Part of Speech) அறிவது அவை எந்த உருபனியல் ஒட்டுகளை ஏற்கும் என்பதை அறியலாம். அவை பனுவலிருந்து பெயர்களையோ முக்கிய சொற்களையோ உதவுவதால் தகவல் மீட்டல் பயன்பாட்டிலும் பயன்படலாம். தானியங்கு சொல்வகைபாட்டு அடையாளப்படுத்திகள் சொல்பொருள் மயக்கம் நீக்கி வழியமைப்பு முறை (Word Sense disambiguation algorithms) சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (Part of Speech – Taggers – Pastaggers) வகுப்பு அடையாளப்படுத்திகள் போன்ற முன்னேறிய ASR மொழிமாதிரிகளில் பயன்படுகிறது.

சொல்வகைப்பாடுகள் தகவல் பிரித்தெடுப்பு பயன்பாடுகள் (Information Extraction Applications)

தமிழக்கான சொல்வகைபாட்டு அடையாளக் குழுமங்கள் (POS) (Parts of Speech tagesets for Tamil)

1. பெயர்
2. மாற்றுப்பெயர்/ பதிலீடுபெயர்
3. அ.ற்றிணைப் பெயர்
4. உயர்திணை பெயர்
5. சுட்டடை
6. பெயரடை
7. எண்ணடை
8. அளவையடை
9. வினையடை
10. காலவினையடை
11. இடவினையடை
12. முறைவினையடை
13. வினை
14. செயபடுபொருள் குன்றிய வினை
15. செயபடுபொருள் குன்றா வினை
16. துணைவினை
17. வினைமுற்று

18. வினை எச்சம்

19. பெயரெச்சம்

20. இநற்தகால வினைமுற்று

21. நிகழ்கால வினைமுற்று

22. எதிர்கால வினைமுற்று

23. இநற்தகாலப் பெயரெச்சம்

24. நிகழ்காலப் பெயரெச்சம்

25. எதிர்காலப் பெயரெச்சம்

26. எதிர்மறைப் பெயரெச்சம்

27. பின் உருபுகள்

28. வேற்றுமை உருபுகள் ஏற்ற பெயர்கள்

ஆங்கிலத்தில் பிரசித்தி பெற்ற POS Taggers Brown Corpus இல் பயன்படுத்தப்பட்ட 87 அடையாளக் குழுமங்களிலிருந்து (Tagsets) உருவாக்கப்பட்டவை.

மூன்று மிகப் பொதுவாகப் பயன்படுத்தப்படுகின்றவை:

1. 45 – Tag penn treebank tagset
2. The Mediurn size 61 tag C5 Tagset used by Lancaster VCREL Projects CLAWS (Constituent Liksihood Automatic word – tagging system) tagger to tag British National Corpus (BNC)
3. The Larger 146 – tag C7 tagset

Ex

Penn tree bank version of brown corpus ASCII flat file

Tag represented after each eord following a slash The/ DT grand /JJ/ jury/NN Commented/ VBD on/ IN a/DT number/ NN of?IN others/ JJ topies/ NNs/

DT - Determiners

JJ - Adjective

NN - Noun, sing, or mass

VBD - Verb, Past tense

IN - Prepossition

NNS - Noun, Plural

பல சொல் வகை அடையாளப்படுத்துவதின் வழிமுறை வரைவு (Algorithm) இரு வகுப்புகளில் அடங்கும். மூன்று Tageers பயன்பாட்டில் உள்ளன.

1. விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Rule – based tageeers)
2. புள்ளியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Statistic Tageers)
3. மாற்றம் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Transformation based Tageers)

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் மயக்கம் தருகின்ற சொல் பெயரா வினையா என்று குறிப்பிடுகின்ற கையால் எழுதிய பெரிய தரவுமையங்களை (Database) உள்ளடக்கியது.

ஆரம்பக்கட்டத்தில் தானாகவே சொல்வகைப்பாட்டை அடையாளப்படுத்த பயன்படுத்தப்பட்ட வழிமுறை வரைவு (Algoritham) இரண்டு நிலை அமைப்பால் ஆனது.

1. அகராதி ஒவ்வொரு சொல்லுக்கும் சாத்தியமான சொல் வகைப்பாட்டைத் தரும்.
2. இரண்டாம் நிலை கையால் எழுதப்பட்ட மயக்கம் நீக்கும் விதிகளை ஒரு சொல்லுக்கு ஒரு சொல் வகைப்பாட்டைத் தருமாறு பயன்படுத்தும்.

ENGTWOL Tagger இரண்டு நிலை அமைப்பால் ஆனது.

ENGTWOL அகராதி இருநிலை உருபனியல் (Two level Morphology) அடிப்படையில் அமைந்தது.

சொல்வகை அடையாளப்படுத்துகை (Parts of Speech Tagging) ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டை அல்லது பிற சொல் வகை அடையாளத்தைத் தருவதாகும்.

அடையாளப்படுத்துகைப் பொதுவான நிறுத்தற்குறிகளுகட்கும் பயன்படுத்தப்படும்.

கணினிமொழியில் அடையாளப்படுத்துகை (Tokenization) போல்தான் இயற்கை மொழியிலும் சொல்லை அடையாளப்படுத்தும் செயல்பாடு அமையும்.

சொல் வகை அடையாளப்படுத்துகை 'பேச்சு புரிந்து கொள்ளுதல்' (Speech recornition) இயற்கை மொழி பகுத்துக்குறித்தல் (Nature Language Parssing) மற்றும் தகவல் மீட்டுப்பெறல் (Information retrivel) என்பனவற்றில் முக்கியமான பங்கு வகிக்கின்றது.

சொல்வகை அடையாளப்படுத்துதலின் உள்ளீடு (Input) வழிமுறை வரைவு (Algoritham) சொற்களின் கோர்வைகளும் குறிப்பிட்ட சொல் வகைப்பாட்டு அடையாளக்

குழுமம் (Part of Speech tagset) ஆகும். வெளியீடு (Output) ஒவ்வொரு சொல்லுக்கும் ஒவ்வொரு நல்ல அடையாளத்தைத் தருவதுதான்.

எ.கா

அந்த <அ.அ> நல்ல <வ.பெ.அ> உயர்நிலைப் <பெ> பள்ளியில் <பெ> படிக்கிறான் <ஒன்றன்பால் நிகழ்கால வினைமுற்று>

எளிய எடுத்துக்காட்டுகளுக்குக் கூட ஒவ்வொரு சொல்லுக்கும் அடையாளம் தருவது அவ்வளவு எளிதல்ல.

எ.கா

அவன் <மாபெ> என் <வி,மா பெ> சொல் <வி.பெ> படி <வி.பெ. பி.உ> கேட்டான் <விழு>/ பெ/ வி/ இ.சொல்/

POS அடையாளப்படுத்தியின் முக்கியமான வேலை சூழலுக்குத் தக்கவாறு சரியான அடையாளத்தைத் தேர்ந்தெடுத்து மயக்கத்தை தீர்ப்பது தான். பெரும்பாலான தமிழ்ச்சொற்கள் எல்லாம் மயக்கம் இல்லாதவை. இருப்பினும் பொதுவாகப் பயன்படுத்தப்படும் சொற்கள் மயக்கம் உள்ளவை.

புள்ளியியல் அடிப்படையிலான POS அடையாளப்படுத்தல் (Statistic Part of Speech Tagging)

சாத்தியங்களை அடையாளப்படுத்துவதின் பயன்படுத்துவது மிகப்பழமையானது பல சாத்திய அடிப்படையிலான அடையாளப்படுத்திகள் வழக்கில் உள்ளன. எ.கா Sidden Markov Moddel or HMM tagget. இந்த புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தியில் பயன்படுத்தப்படும் உள்ளுணர்வு ஒரு சொல்லின் கூடுதலான சாத்தியமான அடையாளத்தை எடு' (Pick up the most likely tag for the word). தரப்பட்ட வாக்கியத்திற்கு அல்லது சொல் கோர்வைக்கு HMM அடையாளப்படுத்தி கோர்வையைத் தேர்ந்தெடுக்கிறது.

P(word/ tag)* P (Tag/ previous n tags)

HMM ஒரு சொல்லைக் காட்டிலும் ஒரு முழு வாக்கியத்திற்கு அடையாளக் கோர்வைத் தெரிந்தெடுக்கும்.

இவ்வகையான Bigram –HMM tager wi என்ற சொல்லுக்கு ti என்ற அடையாளத்தைத் தேர்ந்தெடுக்கும். இது முந்தைய அடையாளம் ti -1 மற்றும் தற்போதைய சொல் wi என்று தரப்படுகையால் கூடுதலான சாத்தியத்திலான அடையாளத்தைத் தேர்ந்தெடுக்கும்.

Ti = argmax p (tj/ti-I,wi)

ஒரு அடையாளப்படுத்தலுக்கு அடிப்படையான HMM சமன்பாடு ti = argmax p(ts/ti –I) p (wi/ti

எ.கா (Brown Corpus)

Secretariat/ NNP is/ VBZ expected/ XBN to/To race/ vB tomorrow/ NN

People/NNS continue/ VBP to/TO inguire/ VB the/DT reason/NN for/IN the/ DT race/NN for/in Outer/ JJ speace/NN

முதல் எடுத்துக்காட்டில் 'race' ஒரு வினை (VB) 2 வது எடுத்துக்காட்டில் 'race' ஒரு பெயர் (NN)

Bigram HMM அடையாளப்படுத்தி அடையாள சிக்கலை அண்மையிலுள்ள சொற்களையும் அடையாளங்களையும் பார்த்துத் தீர்மானிக்கலாம் என்று ஊகம் செய்கிறது.

To/ TO race/???

The/ DT race/???

எவ்வாறு 'race' என்ற நமது எடுத்துக்காட்டு பயன்படுகின்றது என்று பார்ப்போம். நாம் NN –க்கும் VB – க்கும் இடையில் ஒன்றைத் தேர்ந்தெடுக்க வேண்டும். நாம் எதற்குக் கூடுதல் சாத்தியத்தை இருக்கிறதோ அதைத் தெரிந்தெடுக்க வேண்டும்.

P(VB/To) P (race/VB

P (NN/ To P (race/NN)

பின்வரும் சாத்தியங்களாகும் (Probabilities)

P (NN/ To) = :021

P (VB/TO) = .34

மாற்றம் அடிப்படையிலான அடையாளப்படுத்தல் (Transformational based tagging)

மாற்றம் அடிப்படையிலான அடையாளப்படுத்தல் 'Bill tagging' என்று அழைக்கப்படுகிறது. இது பொறிகற்றலுக்கான மாற்றம் அடிப்படையிலான கற்றல் (Transformational – based learning = TBL) அணுகுமுறையாகும். இது விதி அடிப்படையிலிருந்தும் புள்ளியல் அடிப்படையில் இருந்தும் விஷயங்களை எடுக்கிறது. விதி அடிப்படையிலான அடையாளப்படுத்திகளை போல் TBL – உம் எச்சச் சொற்களுக்கு எந்த அடையாளத்தைத் தரவேண்டும் என்று குறிப்பிடுகின்றது. ஆனால் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்திகளைப் போல், TBL ஒரு மொழி கற்றல் நுணுக்கமாகும். இதில் தரவிலிருந்து விதிகள் தானாகவே ஊகிக்கப்படுகின்றது. சில HMM

Taggers – ஐ போல், TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் நுணுக்கமாகும் (Superrised Learning technique), இது ஏற்கனவே அடையாளப்படுத்தப்பட்ட பயிற்சி பெற்ற பொருந்தரவை (Pre – tagged learning corpus) எதிர்பார்க்கின்றது/ ஊகிக்கின்றது.

TBL – உக்கு ஒரு குழுமம் அடையாளபடுத்தும் விதிகள் இருக்கின்றது. முதலில் பெருந்தரவு பெரும்பாலும் பயன்படுத்தப்படும் விரிந்த விதிகளை வைத்து அடையாளப்படுத்தப்படும். பின்னர் சில முன்னதான (Orignal) அடையாளங்களை மாற்றும் மேலும் திட்ட வட்டமான விதிகள் தேர்ந்தெடுக்கப்படும்.

பின்னர் குறைவான எண்ணிக்கையிலான அடையாளங்களை மாற்றும் மேலும் திட்டவட்டமான விதிகள் பயன்படுத்தப்படும். எவ்வாறு TBL விதிகள் பயன்படுத்தப்படுகின்றது (How TBL Rules are Applied)

நாம் Brill's tagger பயன்படுத்தும் ஒரு விதியைப்பார்ப்போம். விதிகள் நடைமுறைப் பயன்படுத்துவதற்கு முன் அடையாளப்படுத்தி ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் (Mostly likely tag) அடையாளப்படுத்தும். இந்த கூடுதல் உகந்ததான அடையாளங்களை நாம் அடையாளப்படுத்தப்பட்ட பொருந்தரவிலிருந்து பெறுவோம்.

எ.கா. Brown corpus – இல் race என்பதற்கு கூடுதல் உகந்தது பெயர் என்பதாகும்.

P (NN/race) = .98

P (VB/ race = .02

நாம் முன்னர் பார்த்த 'race' என்பதன் இரு எடுத்துக்காட்டுகளும் NN என்று அடையாளப்படுத்தப்பட்டிருக்கும். முதல் எடுத்துக்காட்டில் NN என்பது தவறான அடையாளம்.

/S/ VBZ expected/ VBN to/To race/ NN tomorrow/ NN

2 வது எடுத்துக்காட்டில் 'race' என்பது சரியாக NN என்று அடையாளப்படுத்தப்பட்டிருக்கும்.

The/ DT race/NN for/IN outer/JJ space/NN

கூடுதல் உகந்த அடையாளத்தைத் தேர்ந்தெடுத்தபின் Brill's tagger அதன் மாற்ற விதிகளை நடைமுறைப்படுத்தும். இது நடக்கையில் 'rance' என்பதன் தவறான அடையாளப்படுத்தலில் நடைமுறைப்படுத்தப்படும் விதியைக் கற்றிருக்கும். முன்வரும் Tag

To என்றால் NN என்பதை VB என்றும் மாற்றுவதற்கு முன் to/TO என்று வந்தால் இந்த விதி race/NN என்பதை race/VB என்று மாற்றும்.

Expected/ VBN to/TO rrace/NN ----→ expected/ vbn to/To race/VB எவ்வாறு TBL விதிகள் கற்கப்படுகின்றன. (Flow TBl rules are/ carned) Brillis TBL வழிமுறை வரைவுக்கு (algorithm) முன்று முக்கிய நிலைகள் இருக்கின்றன. இது முதலில் ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும். பின்னர் இது ஒவ்வொரு சாத்தியமான மாற்றத்தையும் (Possiable Transformation) பரிசோதிக்கும். கூடுதல் மேம்பாட்டால் அடையாளப்படுத்தலால் விளைந்த ஒன்றை தேர்ந்தெடுக்கும், இறுதியில் இந்த விதிப்படி தரவை மீள் - அடையாளபடுத்தல் செய்யும். முன்னால் வரும் அடையாளப்படுத்தலை விட குறைந்த மேம்பாடுடைய அடையாளப்படுத்தலான நிறுத்தல் அளவீடு வருவது வரை இந்த முன்று நிலைகளும் திரும்பித்திரும்பி நிகழும். 2 – வது நிலை TBL க்கு சரியான ஒவ்வொரு சொல்லின் சரியான அடையாளம் தெரியும் என்பதை வேண்டுகிறது. அதாவது TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் வழிமுறைவரைவு (Supervised Learning algorithm)

TBL செயல்பாட்டின் வெளியீடு (Outbut) முறைப்படுத்தப்பட்ட மாற்றங்களின் பட்டியலாகும். (Ordered list of transformation) இவை புதிய பெருந்தரவுக்கு (Corpus) பயன்படுத்தத்தக்க/நடைமுறைப்படுத்ததக்க அடையாளப்படுத்தல் செயல்முறையை (Tagging Procedure) உருவாக்கும். கொள்கை அடிப்படையில் சாத்தியமான மாற்றங்களின் குழுமம் எல்லையற்றதாகும். ஆனால் மாற்றங்கள் ஒவ்வொன்றும் வழிமுறை வரைவைக் கடக்கும் போது மிகப்பொருத்தமான ஒன்றை தேர்ந்தெடுக்க TBL ஒவ்வொரு மாற்றத்தையும் எடுத்துக்கொள்ள வேண்டும். இவ்வாறு மாற்றங்களின் குழுமத்தை எல்லைப்படுத்த வழிமுறை வரைவுக்கு ஒரு வழி தேவை. இது சிறிய Templats களின் குழுமத்தை (அருமான மாற்றங்கள்) உருவாக்கிச் செய்யப்படுகின்றது. பின்வருவன Brills templates களின் குழுமத்தின் பட்டியலாகும். முன்னர் வருகிற (தொடர்ந்து வருகிற) சொல் Z ஆக அடையாளப்படுத்தப்படும். இரண்டு சொற்களுக்கு முன்னால் (பின்னால்) வருகிற சொல் Z – ஆக அடையாளப்படுத்தப்படும்.

முன்னர் வருகிற இரு சொற்களில் ஒன்று Z ஆக அடையாளப்படுத்தப்படும். முன்னர் வருகிற சொல் Z- ஆகவும் இரு சொற்களுக்கு முன்னால் (பின்னால்) வரும் சொல் W- ஆகவும் அடையாளப்படுத்தப்படும்.

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தல்

முன்னுரை

மக்கள் முன்னேற்றத்திற்கு கணிப்பொறி பல வகைகளில் உதவுகிறது. தொடக்கக் காலங்களில் கணிப்பொறி நல்ல கணக்கீட்டுக் கருவியாக (Calculating Device) பயன்பட்டது. அன்றாட பணிகளிலும், வழக்கமாக செய்ய நேரிடும் பணிகளிலும் கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறியின் வேகம், துல்லியம் செய்திகளை மீளத் தரும் வேகம் மற்றும் நினைவகத் திறன் போன்றன அதிகரித்தன. ஆனால் இதற்கு எதிர் மாறாக விலை குறைந்தது. இதன் விளைவாக கணிப்பொறியின் பயன்பாடு அதிகரித்தது.

ஆக்கப்பூர்வமான (Creative) பணிகளுக்கும் கணிப்பொறிகள் பயன்படும் வண்ணம் உருவாக்கப்படுகின்றன. குறிப்பாக, கல்விக்கூடங்களில் பயிற்றுவிப்பதற்கு பயன்படும் கருவியாக பயன்படுத்தப்படுகின்றன. கணிப்பொறி வழி பாடங்களை தயாரித்து உரிய திட்டங்களை செயல்படுத்தி மாணவர்களுக்கு கற்பிக்கலாம். கணிப்பொறி பாடங்களை கற்பதற்கும் கற்பிப்பதற்கும் எங்ஙனம் உதவுகிறது என்பது பற்றி இங்கு காண்போம்.

கற்றல் கற்பித்தலில் கணிப்பொறிகள்

அடிப்படையில் கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. தேவையான பணிக்கான வழியமைப்பு அல்லது மென்பொருள் கொண்டு அது இயக்கப்படுகிறது. இருப்பினும் செய்திகளை சேகரித்தல், ஆராய்தல், வெளியிடுதல் மற்றும் பரப்புதல் போன்ற செயல்பாடுகளில் மிக மிக அதிவேகமுடையது. மேலும் கணிப்பொறியில் உரைச்செய்திகள், வரைபடங்கள், புகைப்படங்கள், ஒலிச்செய்திகள் மற்றும் சலனப்படங்கள் (ANIMATION) போன்ற பல வகையான செய்திகளை உள்ளீடு செய்து ஆராய்ந்து முடிவு பெறுவதற்கான வசதிகள் உள்ளன. எல்லா வகைச் செய்திகளையும் கலந்த கலவையாகவும் செய்திகள் அமையப்பெறலாம். இது போன்ற எல்லா வகை செய்திகளையும் கலந்த கலவையை பல்லூடகம் (Multimedia) என்கிறோம். மற்றொரு பக்கத்தில் பயிற்றுவிப்போர் அல்லது ஆசிரியருக்கு இருக்க வேண்டிய அடிப்படை இயல்புகள் மூன்று ஆகும்.

அவைகளாவன

1. செய்திகளை திரட்டுதல்
2. செய்திகளை ஆராய்தல்
3. பயில்வோருக்கு ஏற்றவாறு செய்திகளை எடுத்துரைத்தல்

இத்தகைய குணங்களை கணிப்பொறி பெற்றிருப்பதால் கற்பித்தலுக்கும், கற்பதற்கும் சிறந்த சாதனமாக பயன்படுத்தப்படுகிறது. கணிப்பொறியால் சொந்தமாக பாடங்களை உருவாக்க இயலாது. இவ்வாறான கணிப்பொறியை ஒரு ஆசிரியரைப் போன்று செயல்பட வைப்பதற்கு மென்பொருள்கள் (Software) உருவாக்கப்படுகின்றன. ஒரு பாடத்தை பயிற்றுவிப்பதற்காக உருவாக்கப்படும் மென்பொருள் பொதுவாக (Course ware) எனப்படுகிறது. கணிப்பொறி வழி கற்பித்தலுக்கு CAI (Computer Assisted/Aided Instruction) என்று பெயர். Assisted or Aided என்ற சொல் கற்பித்தலுக்கு கணிப்பொறி உதவும் என்ற கருத்தை வலியுறுத்தும். அதுபோல் கணிப்பொறி மூலம் பயிலும் முறைக்கு CAL (Computer or Assisted /Aided Learning) என்று பெயர். CAL and CAI போன்ற சொற்களுக்கு நிகராக பல தொடர்கள் பயன்படுத்தப்படுகின்றன. CMI- (Computer Managed Instruction), CMI – Computer Managed Learning, CDI – Computer Directed Instruction Learning போன்றன சிலவாகும். இத்தொடர்களில் Managed and Directed என்ற சொற்கள் கற்பித்தலில் கணிப்பொறிக்கு உள்ள முக்கியத்துவத்தை பிரதிபலிக்கும். இத்தகைய அமைப்புகளால் உருவாக்கப்படும் மென்பொருள்கள் பாடங்களை பயிற்றுவிப்பதோடு பயில்வோரின் திறன்களை ஆராய்கிறது. கற்பிக்கின்ற பொழுது பயில்வோரிடம் கேள்வி கேட்கிறது. பயில்வோரின் விடைகளை சரியான விடையோடு ஒப்பிட்டு நோக்குகிறது. மதிப்பீடு செய்கிறது. மதிப்பீட்டின் அடிப்படையில் பயில்வோரை வழிநடத்துகிறது.

மொழியினை கற்பிப்பதற்காக உருவாக்கப்படும் மென்பொருள்கள் CALT- Computer Aided Language Teaching என்று பெயர். அதுபோல் மொழிகள் படிப்பதற்காக உதவும் மென்பொருள்கள் CALL – Computer Aided Language Learning என்று பெயர்.

கணிப்பொறி வழி கற்றல் கற்பித்தல் - சிறப்பியல்புகள்

கணிப்பொறி வழி கற்றல் கற்பித்தலில் எண்ணற்ற பயன்களும் சிறப்பியல்புகளும் உள்ளன. பொதுவாக சிறப்பியல்புகளை முன்று வகைகளாகப் பிரிக்கலாம். அவைகளாவன.

1. கணிப்பொறியில் இயல்பாக இருக்கின்ற சிறப்பு குணங்கள்
2. பயில்வோருக்கு கிடைக்கின்ற பயன்கள்
3. பயில்வோருக்கு உதவும் சிறப்புகள்

கணிப்பொறியின் செயல், வேகம், துல்லியம், சோர்வின்மை, சகிப்புத்தன்மை (Talarance) நெகிழ்வுத்தன்மை ஒரே நேரத்தில் பல செயல்களை புரியும் திறமை போன்றன கணிப்பொறியில் இயல்பாக அமைந்துள்ள பண்புநலன்கள் ஆகும். கணிப்பொறிகள் மனிதனைப்

போல் சோர்வடைவது இல்லை. எனவே எப்பொழுது வேண்டுமானாலும் எவ்வளவு நேரம் வேண்டுமானாலும் பயன்படுத்தலாம். எனவே கணிப்பொறிகள் கல்வி கற்றல் கற்பித்தலில் பயன்படுத்தப்பட்டு வருகின்றன. மேலும் வேறு எந்த ஒரு கருவியிலும் இல்லாத பல்லூடக (Multimedia) சாதன அமைப்புக் கொண்டது. எனவே உரைநடைச் செய்திகள் மட்டுமல்லாது வரைபடங்கள், புகைபடங்கள் மற்றும் ஒலிசெய்திகள் போன்றவைகளை ஆய்வதற்கும், கையாள்வதற்கும் எண்ணற்ற வசதிகளைப் பெற்றுள்ளது. இதனால் கல்வி கற்பிப்பதற்கும் கற்பதற்கும் தேவையான அனைத்து அடிப்படை வசதிகளையும் பெற்றுள்ளது. படவில்லைகள் (Slides) புகைப்படங்கள் மற்றும் கல்வித் தொடர்பான சினிமா படங்கள் கற்றலுக்கும் கற்பித்தலுக்கும் சிறந்த சாதனங்களாக பயன்படுத்தப்பட்டு வருகின்றன. இவைகளை கருவியாக பயன்படுத்திக் கொள்ள இயலுமெனினும் அந்த நிகழ்ச்சிகளுடன் உரையாடுவதற்கோ, கற்பிப்பவரின் தன்மைக்கேற்ப கட்டுப்படுத்தவோ இயலாது. அந்த நிகழ்ச்சிகளை தயாரித்த வரிசை கிரம முறையில் மட்டுமே பார்க்கலாம். கற்போரின் மனநிலைக்கேற்ப தேவையான செய்திகளை கூடுதலாகவோ அல்லது குறைவாக பெறவோ இயலாது. கற்போரின் ஐயப்பாடுகளை களைவதற்கு தகுந்த வண்ணம் அந்த நிகழ்ச்சிகளுடன் தொடாட்பு கொள்ளவோ இயலாது. ஆனால் கணிப்பொறி வழி இத்தகைய கல்வி நிகழ்ச்சிகளை அல்லது பாடங்களை பதிவு செய்யும் பொழுது அதனுடன் உரையாட இயலும் தேவையான செய்திகளை தேவையான வண்ணம் கேட்டுப் பெறலாம். இது கணிப்பொறி கருவியின் மிகப் பெரிய சிறப்பம்சம் ஆகும். கணிப்பொறியை எந்த நேரமும் எப்பொழுதும் பயன்படுத்தலாம் என்பதால் பயில்வோர் விருப்பமான நேரத்தில் விருப்பமான பாடத்தை பயிலலாம். கணிப்பொறி வழி பயில்வோரின் சுதந்திரம் தேவை. இங்கு பயில்வோர் தனித்தனியாக பயில வாய்ப்புகள் உண்டு.

1. பயில்வோரின் ஈடுபாடு
2. கற்கும் திறன்
3. கற்கும் வேகம்

ஆகிய தன்மைகளைப் பொறுத்து பாடங்களைக் கற்றுக் கொள்ளலாம். சாதாரண வகுப்புகளில் இத்தகைய சாத்தியங்கள் இல்லை. பாடங்களை விரைவாக கற்றுக் கொள்பவர்கள் மெதுவாக கற்றுக் கொள்பவரின் வேகத்திற்கு ஒத்துப் போக வேண்டும். இதனால் வேகமாக கற்றுக் கொள்பவரின் வேகம் வீணாகிப் போகும். வகுப்பில் கூச்ச சுபாவத்தால் ஐயப்பாடுகளை களைவதற்கான கேள்விகளைக் கேட்பதற்கு மாணவர்கள் தயங்குவர். ஆனால் கணிப்பொறியில் சுயமாக தனியாக படிக்கின்ற பொழுது எவ்வித கூச்சமும் தயக்கமும் இன்றி ஐயப்பாடுகளை களைய விளக்கம் பெற இயலும். மேலும் கணிப்பொறி வழி கற்கின்ற பொழுது மனநிலையைப்

பொறுத்து ஒரு பாடத்தில் அதிக செய்தியையோ முக்கிய செய்தியையோ படித்து அறியலாம். இதுபோன்று ஒரு பாடத்தை படித்தறிவதற்கான பரிமாணங்களை அல்லது சிறப்புகளை பயில்வோரே தீர்மானித்துக் கொள்ளலாம். மேற்கூறியவைகள் எல்லாம் கணிப்பொறி வழி பயில்வோருக்கு உள்ள சிறப்புத் தன்மைகளாகும்.

கணிப்பொறி வழி கற்பிப்பதற்கான பாடங்களை உருவாக்கும் ஆசிரியர்கள் மாணவர்களின் கல்வி நிலைக்கேற்ப கற்கும் திறனுக்கேற்ப வடிவமைக்க வேண்டும் என்பதில்லை. ஒரு பொதுவான பாடங்களை தயாரித்து பயில்வோரின் திறமைக்கேற்ப அவைகளை உருவாக்கித் தர இயலும். உரைநடைச் செய்திகள், விளக்கப்படங்கள் அனைத்தும் இதில் உள்ளதால் பாடக்கருத்துக்களை எளிதில் புரிய வைக்க இயலும். ஒரு முறை பாடத்திட்டங்களை வகுத்து செம்மையாக தயாரித்து கணிப்பொறியில் ஏற்றி எத்தனை முறை வேண்டுமானாலும் பயன்படுத்தலாம். ஆயிரங்கணக்கானவர்களுக்கு இத்தகைய பாடங்கள் பயன்படுத்துவதற்கு சிறப்பாக அமையும். இத்தகைய பாடங்களை ஒரு துறை நிபுனத்துவம் பெற்ற ஆசிரியர்கள் உருவாக்குகின்றனர். அவர்களின் திறமை உலகம் முழுவதும் உள்ள ஆயிரங்கணக்கானவர்களுக்குப் பயன்படுத்தலாம்.

CAI – உருவாக்குவதற்கான கணிப்பொறிமொழி

CAI கணிப்பொறி வழி கற்பித்தல் மற்றும் கற்றல் பயன்பாடுகளுக்கான மென்பொருள்களை உருவாக்குவதற்கென்றே கணிப்பொறி அமைப்பு மொழிகள் உள்ளன. IBM – International Business Machine என்கின்ற நிறுவனம் Course Writer என்ற வழியமைப்பு மொழியை உருவாக்கியுள்ளது. இதுபோல் PILOT என்ற மொழியை Control Data Corporation என்ற நிறுவனமும் TW என்ற மொழியை Stark Weather என்ற நிறுவனமும் Extol என்ற மொழியை Kenning and Kenning என்ற நிறுவனமும் தயாரித்து வெளியிட்டு உள்ளன.

கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான பாடங்களை உருவாக்க மேலும் கணிப்பொறி மொழிகள் உதவுகின்றன. இம்மொழிகள் ஏனைய கணிப்பொறி மொழிகளை விட எளிமையானவை. இருப்பினும் ஒரு கணிப்பொறி மொழி கற்பதற்கான அதைப் பயன்படுத்துவதற்கான திறன் பெற்றவர்கள் மட்டுமே இவைகளைப் பயன்படுத்த இயலும். எனவே இவைகளைக் கொண்டு பாடங்களை உருவாக்குபவர்கள் இக்கணிப்பொறி மொழியில் தேர்ச்சிப் பெறுவது அவசியமாகும். மேலும் வழியமைப்பு திறன் கொண்டவர்களாகவும் இருத்தல் வேண்டும்.

CAI உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள் (Packages)

பொதுவாக வழியமைப்பு மொழிகளைப் பயன்படுத்தி ஒரு திட்டங்களை உருவாக்குகின்ற போது திட்டங்களை செயல்படுத்துவதற்கான முறைமைகளில் தேர்ச்சிப் பெற்றிருக்க வேண்டும். உதாரணமாக பத்து சொற்களை கணிப்பொறிக்கு கொண்டு அகர வரிசை படுத்துவதாகக் கொள்வோம். இப்பணியை வழியமைப்பு மொழியை பயன்படுத்தி செய்ய வேண்டுமெனில் வழியமைப்பு மொழியில் தேர்ச்சி மட்டும் போதாது. அகர வரிசைப்படுத்துவதற்கான முறைகளை தெரிந்தவராகவும் இருத்தல் வேண்டும். அவ்வாறு இருந்தால் தான் அதன் முறைகளையும் செயல்பாடுகளையும் கணிப்பொறிக்கு கட்டளைகளாக தந்து செயல்படுத்த இயலும். ஆனால் பயனீட்டு மென்பொருள்களில் இதுபோன்ற செயல்பாடுகளுக்கு தேவையான முறைகள் கணிப்பொறியாக்கம் செய்யப்பட்டு தயார் நிலையில் உள்ளன. அவைகளை பயன்படுத்திக் கொள்வதற்கான சாதாரண திறமைகள் இருந்தால் போதுமானது. எனவே கணிப்பொறி வழி கற்றல் கற்பிப்பதற்கான மென்பொருள் (கனிமம்) உருவாக்க கணிப்பொறி மொழிகளைப் பயன்படுத்த முனைவோர் பாடத்திட்டங்களை மட்டும் அல்லாது கணிப்பொறி மொழியிலும் தேர்ச்சி பெற்றவராக இருத்தல் வேண்டும். இது ஒரு மிகப் பெரும் தடையாகும். இத்தடைகளை அகற்றவும், இப்பணிகளை எளிமைப்படுத்தவும் பல பயனீட்டு மென்பொருள்கள் (Utility Software) வெளிவருகின்றன. இவைகளில் பாடங்களை கணிப்பொறி வழி தயாரிப்பதற்கான அனைத்து வசதிகளும் தயார் நிலையில் உள்ளன. இத்தகைய பயனீட்டு மென்பொருள்களை தயார் நிலையில் உள்ள வசதிகளை எப்படி பயன்படுத்திக் கொள்வது என்ற பயிற்சி மட்டுமே போதுமானது. இவைகளை எளிதாக கற்றுக் கொள்ளலாம். கணிப்பொறி பயிற்சி இல்லாதவர்கள் எளிமையாகவும் விரைவாகவும் கற்றுக் கொள்ள வசதி உள்ளது. Story Board மற்றும் Copywriter போன்றன. CAI உருவாக்குவதற்கு பயன்படுத்தப்படும் பயனீட்டு மென்பொருள்கள் ஆகும். இவைகள் தவிர light Sources என்ற மென்பொருளும் உள்ளது.

CAI இல் நகரும் படங்கள் (Animation)

CAI மென்பொருள் உருவாக்குவதில் நகரும் படங்கள் அல்லது சலனப்படங்கள் மற்றும் வீடியோ காட்சிகள் போன்றவைகளை அமைப்பது மென்பொருளின் தரத்தை மேம்படுத்தும். இதற்காக சில மென்பொருள்கள் உள்ளன. அவற்றுள் Video Works என்பது ஒன்றாகும். இம்மென்பொருளில் சலனப்படங்களை உருவாக்கவும் தேவையான வசதிகள் உள்ளன.

மனித நுண்ணறிவின் பங்கு (Artificial Intelligence)

CAI உருவாக்கத்தில் கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. மனிதனைப் போன்று சிந்தித்து செயல்படக்கூடிய கணிப்பொறிகளை உருவாக்க அறிஞர்கள்

முனைந்துள்ளனர். கணிப்பொறியை மனிதரைப்போல் உருவாக்கி சூழலுக்கேற்ப பணியாற்றவும் அனுபவத்தில் புதிய செய்திகளை சுயமாக கற்கவும் தேவையான திறமைகளை உருவாக்குகின்றன. இத்துறைக்கு செயற்கை நுண்ணறிவு (Artificial Intelligence) என்று பெயர். இதை AI என்று சுருக்கமாக அழைப்பர். இத்துறையின் அங்கமாக இயற்கை மொழி ஆய்வு NLP (Natural Language Processing) என்ற துறை உள்ளது. ஒரு மனிதன் மற்றவர்களிடம் உரையாடுவது போல் மனித மொழியில் கணிப்பொறியுடன் உரையாட தேவையான வசதிகளை இத்துறை செய்கிறது. இந்த ஆய்வில் வெற்றி பெறுவேமானால் கணிப்பொறிகளுடன் உரையாட அல்லது கணிப்பொறிக்கு கட்டளையிட கணிப்பொறி மொழி என்ற ஒன்று தேவையில்லை. மனித மொழியில் கணிப்பொறிக்கு கட்டளை தரலாம். அதை கணிப்பொறி புரிந்து கொண்டு செயலாற்றும். அதுபோல் கணிப்பொறியின் விடையை மனித மொழியில் விளக்கமாக தரும். இத்தகைய அமைப்புகள் மனிதனுக்கும் கணிப்பொறிக்கும் இருக்கின்ற உறவை எளிமைப்படுத்தும். மேலும் துறை ஒருங்கு (Expert System) ES என்பது கணிப்பொறி எழுதியுள்ள மிகப்பெரும் பிரிவாகும். கணிப்பொறியை ஒரு துறையின் வல்லுநரைப் போன்று பயன்படுத்த இவ்வாய்வுகள் உதவும்.

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்காக மென்பொருள்களை உருவாக்கும் பொழுது மேற்கூறிய தொழில் நுட்பங்கள் (மனித நுண்ணறிவு, இயற்கை மொழி ஆய்வு, துறை ஒருங்கு) போன்றன சேர்க்கப்படுகின்றன. செயற்கை நுண்ணறிவுடன் கூடிய CAI – மென்பொருளை ACAI என்று அழைக்கிறார்கள். இதை ITS (Intelligence Tutorial System) என்றும் அழைப்பர்.

கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் கணிப்பொறி சிறந்த பங்காற்றுகிறது. ஆனால் மொழி கற்றல் கற்பித்தல் போன்ற செயல்களில் எண்ணற்ற நுணுக்கங்கள் தேவைப்படுகின்றன. இதில் கணிப்பொறி ஓரளவுதான் பங்காற்ற முடியுமே தவிர பயிற்றுவிக்கும் ஆசிரியருக்கு நிகராக அமையாது. மொழியை கற்பிப்பதற்கு கணிப்பொறி துணைக்கருவியாக மட்டுமே பயன்படும். இருப்பினும் இத்தகைய மென்பொருளை உருவாக்குகின்ற பொழுது எது இயலுமோ அதில் உறுதியும் திடமான சிந்தனையும் வேண்டும். மொழி கற்றல் கற்பித்தலில் கணிப்பொறியில் செய்யத்தகும் பயிற்சிகள் கீழே பட்டியலிடப்படுகின்றன.

1. கணிப்பொறி வழி உரையாட பயிற்சி அளித்தல்
2. மொழியின் கோடிட்ட இடங்களை நிரப்புதல்
3. ஒரு சொல்லைத் தந்து அதற்கான உச்சரிப்பு

4. வாக்கிய வகைகளை மாற்றி அமைத்தல்
5. இரு சிறு வாக்கியங்களை இணைத்து ஒரு வாக்கியமாக உருவாக்குதல்
6. சொல்லின் ஒருமை, பன்மை, எதிர்பதங்கள் போன்றவற்றை கற்றுக் கொள்வதற்கான பயிற்ச்சிகள்
7. சொற்களை பிரித்து எழுதுதல்
8. சொற்களை சேர்த்து எழுதுதல்
9. ஒரு சொல்லின் இலக்கண கூறினை கொடுத்து மற்ற இலக்கண கூறுகளை கேட்டல்

மேற்கூறிய பட்டியலில் கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான சில வாய்ப்புகள் கொடுக்கப்பட்டுள்ளன. இது போன்று எளிமையான பயிற்ச்சிகளை தருவதற்கு மென்பொருள்களை உருவாக்கி பயன்படுத்தலாம்.

கணிப்பொறி வழி கற்றல், கற்பித்தலுக்கான மென்பொருள்கள்

கணிப்பொறி வழி கற்பித்தலுக்கு எண்ணற்ற மென்பொருள்கள் உள்ளன. அறிவியல், தொழில்நுட்பம் போன்ற பாடங்களை பயிற்றுவிப்பதற்கு உலகளவில் பல்லாயிரக்கணக்கான மென்பொருள்கள் உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன.

அதுபோல் மொழி கற்றல் கற்பித்தலுக்கு பல மென்பொருள்கள் உள்ளன. கணிப்பொறி வழி தமிழ் கற்றல் கற்பித்தலுக்கு மென்பொருள்கள் உருவாக்கம் தொடங்கப்பட்டுள்ளது.

சென்னையைச் சார்ந்த Learn Fun என்ற நிறுவனம் தமிழுக்காக மென்பொருள் ஒன்றை உருவாக்கியுள்ளது. இம்மென்பொருளில் தமிழ் எழுத்துக்களை எழுத படிக்க வசதிகள் உள்ளன. இம்மென்பொருள்கள் மொழி விளையாட்டு மென்பொருள் வகையைச் சேர்ந்தது. தமிழை ஆரம்ப நிலையில் கற்பவர்கள் வல்லினம், மெல்லினம், குறில், நெடில் போன்ற அடிப்படைக் கூறுகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள்கள் மிகவும் உதவுகிறது.

சென்னையைச் சார்ந்த Soft View என்ற நிறுவனம் அமுதம் என்ற மென்பொருளை உருவாக்கி உள்ளது. இம்மென்பொருள் திருக்குறள், ஆத்திச்சூடி போன்றவைகளை எளிமையாக கற்றுத் தருகிறது.

சிங்கப்பூர் தேசிய கல்விக் கழகம் தமிழில் பயிற்றுவிப்பதற்கான மென்பொருள் ஒன்றை உருவாக்கியுள்ளது. ஒருமை, பன்மை, எதிர்பதம் போன்றவைகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள்கள் உதவுகிறது.

சிங்கப்பூரைச் சார்ந்த தமிழாசிரியர் கழகம் தமிழைப் பயிற்றுவிப்பதற்காக சில மென்பொருளை வெளியிட்டுள்ளன. அவைகளைக் கொண்டு தமிழில் உள்ள இரண்டு எழுத்து

சொற்கள் முன்று எழுத்து சொற்கள் போன்றவைகளை எளிமையாக கற்பிக்கலாம். மேலும் வெளிப்படும் படத்திலிருந்து அப்படத்தில் உள்ள பொருளை அடையாளம் கண்டுப் பயிற்ச்சி பெற சில வசதிகள் இம்மென்பொருளில் உள்ளன. மேலும் ஒன்றிலிருந்து ஆயிரம் வரை எண்களை எழுதவும், படிக்கவும் சொல்லித் தருகிறது. மேற்கண்ட மென்பொருள் தயாரிக்கும் பணி அமரர் பேராசிரியர் கோவிந்தசாமி தலைமையில சிங்கப்பூரில் நடைபெற்றது.

அமெரிக்காவை சார்ந்த விண்டஹோம் என்பவர் தமிழ்ச் சொற்கோவைகளை பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை வெளியிட்டுள்ளார். பென்சில்வேனியா பல்கலைக்கழகத்தில் பேச்சுத் தமிழைப் பயிற்றுவிப்பதற்காக மென்பொருள் உருவாக்க முனைந்துள்ளனர். அமெரிக்காவில் உள்ள சிகாகோ நகரிலுள்ள நிறுவனம் ஒன்று Visual Tamil என்ற மென்பொருளை உருவாக்கியுள்ளது. இம்மென்பொருள் தமிழ் கலைப்பண்பாடு ஆகியவற்றை கற்பிக்கிறது. இது போன்று ஆங்காங்கே கணிப்பொறி வழி தமிழில் கற்பித்தல் கற்றலுக்கு பல மென்பொருள்கள் உருவாக்கப்படுகின்றன.

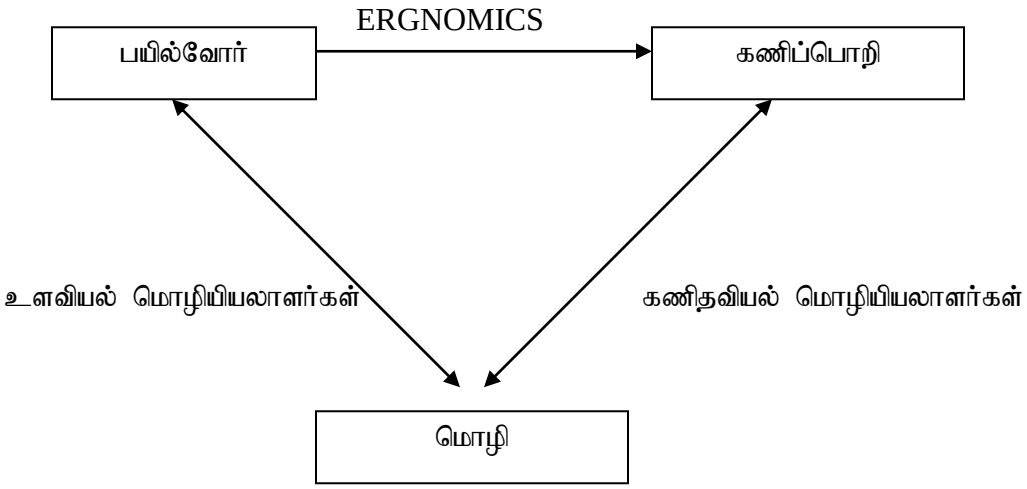
கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு:

ஒன்றுக்கு மேற்பட்ட கணிப்பொறி ஒன்றோடு ஒன்று இணைக்கப்படுகின்றன. இத்தகைய அமைப்பிற்கு கணிப்பொறி வலைப்பின்னல் (Computer Networks) என்று பெயர். இதனால் ஒரு கணிப்பொறியில் பதிவு செய்யப்பட்டுள்ள மென்பொருள் மற்றும் பதிவுகளை பிற கணிப்பொறிகளில் பெறலாம். இது போன்று உலகலாவிய நிலையில் உள்ள கணிப்பொறிகளை ஒன்றிணைத்து தகவல் பரிமாற்றம் செய்துகொள்ள உதவும் அமைப்பிற்கு இணையம் என்று பெயர். இதன் பயனாய் இணையத்தில் உள்ள ஒரு கணிப்பொறியில் பதிவு செய்யப்படும் செய்திகளை உலகில் உள்ள எந்த கணிப்பொறியிலிருந்தும் பெறலாம். கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான வழிமுறைகள் இணைய ஏற்றம் செய்யப்பட்டுள்ளன. இம்மென்பொருள்களை செயல்படுத்தி மொழிகளைக் கற்பிக்கலாம். அல்லது கற்கலாம். பென்சில்வேனியா பல்கலைக்கழகத்தில் WALTT Web Aided Language Teaching and Training] என்ற திட்டம் செயல்படுத்தப்பட்டு உள்ளது. இணையத்தின் வழி பேச்சுத் தமிழைப் பயிற்றுவிப்பதற்கு இது உதவுகிறது. மேலும் தமிழ் மொழியை கற்பிப்பதற்கு சிங்கப்பூரில் உருவாக்கப்பட்ட பல இணையதளங்கள் செம்மையாக உள்ளன. தமிழ்ப்பல்கலைக்கழகத்தின் ஒரு அங்கமாக தமிழ் இணையப் பல்கலைக்கழகம் செயல்படுகிறது. இணையப் பல்கலைக்கழகம் தமிழ்ப்பாடத்தில் சான்றிதழ் வகுப்புகளையும் பட்டைய வகுப்புகளையும் தொலைதூர கல்வி திட்டத்தின் கீழ் இணையம் வழி நடத்துகிறது. அயல்நாட்டினர் அயல்நாட்டில் உள்ள தமிழர்கள் மற்றும் இந்திய மாநிலங்களில் உள்ள தமிழர்கள் தமிழ்

மொழியை இதன் வழி கற்கலாம். இப்பல்கலைக்கழகம் தமிழ்மொழி பண்பாடு, கலாச்சாரம் போன்றவற்றை இணையம் வழி கற்பிக்கிறது. மேலும் தமிழ்மொழியை கற்றுக்கொள்வதால் உறுதுணையாக அமையும். அகராதிகள், இலக்கியங்கள் போன்றவைகளை இந்த இணையத்தின் முலம் பெறலாம். இது பேன்று தமிழ் மொழி கற்றலுக்கும், கற்பித்தலுக்கும் இணையத்தில் பல வாய்ப்புகள் உள்ளன.

மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு:

இத்தகைய மென்பொருளை உருவாக்க மொழி வல்லுநர் உளவியல் மொழியியளர்கள் (Psycho linguistics) கணித மொழியியல் வல்லுநர்கள் (Computer Linguistics) ஆகியோரின் ஒத்துழைப்பு அவசியம் ஆகும். மொழியியல் வல்லுநர்கள் கணிப்பொறிக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். உளவியல் மொழியியலாளர்கள் பயில்வோருக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். மேலும் கணிப்பொறிக்கும், பயில்வோருக்கும் உள்ள உறவை (ERGNOMICS) என்ற துறையினர் ஆராய்கின்றனர். இத்தொடர்பு கீழ்கண்ட கட்டப்படம் விளக்குகிறது.



நிறைவு :

கணிப்பொறி வழி கற்றல், கற்பித்தலில் மேற்கூறியவாறு கணிப்பொறி பல வகைகளில் பயன்படுகிறது.

கணிப்பொறி வழி அகராதி தயாரித்தல்:

மொழி கற்றலில் அகராதியின் பங்கு மகத்தானது. ஒரு மொழியின் சொல் அதன் இலக்கணக்கூறு, பொருள், பயன்பாடு மற்றும் வாக்கியத்தில் அதன் உபயோகம் போன்றவைகளை அகராதி விளக்குகிறது. மேலும் ஒரு சொல்லுக்கு நிகரான பிற மொழி சொற்களையும் இரு மொழி அகராதிகள் விளக்குகின்றன. அகராதி தயாரிக்கும் பணி

எளிதானதல்ல. மிகுந்த நேரமும் அதிக உடல் உழைப்பும், பொருட்செலவும் ஒரு அகராதியை உருவாக்க நீண்ட காலம் தேவைப்படும். எனவே அகராதியை உருவாக்க கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன. இது பற்றி சுருக்கமாக இங்கு காண்போம்.

அகராதி உருவாக்கும் படி நிலைகள்:

பொதுவாக அகராதி ஒன்றைத் தயாரிக்கத் தேவைப்படும் படிநிலைகளை இரண்டு வகையாகப் பிரிக்கலாம்.

1. செய்திகளை சேகரித்தல் ஆராய்ந்து வகைப்படுத்துதல், அகரவரிசைப்படுத்துதல் இப்படிநிலைகள் முதற்கட்ட நிலைகளாகும்.
2. சொல்லுக்கான இலக்கணகூறு ஆராய்ந்து பொருள் எழுதுதல், பிறமொழி சொற்களை எழுதுதல், சொல்லின் உபயோகத்தை விளக்குதல் இவைகள் இரண்டாம் கட்ட நிலை பணிகள் ஆகும்.

இவ்விரு கட்டப்பணிகளில் முதல் கட்டப் பணிகளை எளிமையாக கணிப்பொறிப்படுத்த இயலும். இரண்டாவது நிலை கட்டப் பணிகளை மனித முயற்சியிலேயே செய்ய இயலும்.

கணிப்பொறி வழி அகராதி – வரலாற்றுச் சுருக்கம்:

மார்த்தூஸ் அர்னால்ட் (Mathews Arnald) என்பவர் ஒரு ஆங்கில கவிஞர். இவரது கவிதைகளுக்கான சொல்லடைவு முதன்முதலாக கணிப்பொறி வழி உருவாக்கப்பட்டது. இதற்கு IBM -704 என்ற கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறி வழி இச்சொல் அடைவினை முதன்முதலாக ஸ்டீபன் (Stephen) உருவாக்கினார். இதைத் தொடர்ந்து Oxford அகராதி உருவாக்கத் தேவையான சொல்லடைவு (Concordance) கணிப்பொறி வழி உருவாக்கப்பட்டது. இதைத் தொடர்ந்து ஆங்கிலம் மற்றும் பிற மொழிகளின் அகராதி உருவாக்க கணிப்பொறிகள் ஆங்காங்கே உருவாக்கப்பட்டன. தமிழ்ப் பல்கலைக்கழகத்தில் சங்க இலக்கியத்திற்கான அகராதி உருவாக்கக் கணிப்பொறி பயன்படுத்தப்பட்டது. தமிழ் அகராதி உருவாக்க கணிப்பொறி பயன்படுத்தப்பட்டது. இதுவே முதல் முயற்சியாகும். கிரியா என்னும் தற்கால அகராதி கணிப்பொறி வழி உருவாக்கப்பட்டது. இந்த அகராதி அச்சு வடிவிலும் குறுந்தகடுகளிலும் கிடைக்கப்பெறுகின்றன.

கணிப்பொறி வழி அகராதி – முறைமைகள்

கணிப்பொறி வழி அகராதி உருவாவதற்கு கீழ்கண்ட முறைமைகள் பின்பற்றப்படுகின்றன.

1. செய்திகளை உள்ளீடு செய்தல்

2. செய்திகளை வார்த்தைகளாகப் பிரித்தல். இந்நிலையில் ஒரு வாக்கியத்திலிருந்து சொற்கள் தனியாகப் பிரிக்கப்பட்டு பின்வரும் பட்டியல் உருவாக்கப்படுகிறது.

சொல் Word	வரிசைஎண் Line Number

3. மேற்கண்ட பட்டியலில் சொற்கள் அவைகள் இடம்பெறும் வரிசையில் இடம்பெறுகின்றன. மேலும் ஒரு சொல் உரையின் பல இடங்களில் பயன்படுத்தப்பட்டிருந்தால் அத்தனை முறை அச்சொல் இப்பட்டியலில் இடம்பெறும்.
4. மேற்கண்ட பட்டியல் அகரவரிசைப்படுத்தப்படுகிறது. இது மேற்கண்ட பட்டியலில் சொற்கள் இடம்பெறும் வரிசைமுறையை மாற்றுகிறது. ஒரு சொல் பலமுறை இடம்பெற்றிருப்பினும் அது அடுத்தடுத்து வரிசைமுறை பெறுகிறது.
5. இவ்வாறு அகர வரிசைப்படுத்தப்பட்டுள்ள பட்டியலிலிருந்து கீழ்கண்ட பட்டியல் உருவாக்கப்படுகின்றன.

சொல் Word	நிகழ்வெண்ணிக்கை Frequency	சொல் இடம்பெறும் இடம்

இப்பட்டியலின் உதவிக் கொண்டு பின்வரும் பட்டியல் உருவாக்கப்படுகிறது.

சொல் Word	நிகழ்வெண்ணிக்கை Frequency	சொல் இடம்பெறும் வரி எண் List of Line Number	இடம்பெறும் வரி எண் Citation

மேற்கண்ட பட்டியலிலிருந்து ஒரு சொல் உரையில் எத்தனை முறை பயன்படுத்தப்பட்டுள்ளது என்பதை அறியலாம். அச்சொல் இடம்பெறும் வரி எண்கள் மற்றும் வரிகள் தயார் நிலையில் கிடைக்கப்பெறுகின்றன. சொல் இடம்பெறும் வரி தெரிவால் அதனைக் கொண்டு அச்சொல்லின் இலக்கணக்கூறு, பொருள் உபயோகம் போன்றவைகளை ஆராய்ந்து எழுதலாம். இக்கட்டநிலைப் பணிகளை கணிப்பொறியை அகராதி தயாரிப்பதற்கு செம்மையாக பயன்படுத்திக்கொள்ளலாம். விபரங்கள் துல்லியமாகவும் விரைவாகவும் கிடைக்கப்பெறும்.

அகராதி தயாரித்தலில் இயற்கை மொழி ஆய்வு:

கணிப்பொறி வழி அகராதி தயாரிக்கப்படும்போது கிடைக்கப்பெறும் இறுதிபடிவம் சொல்லடைவுப் படிவம் போன்று இருக்கும் இதில் பெயர்ச்சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர்ச்சொற்கள் தனித்தனி சொற்களாகவே இடம்பெறும். உதாரணமாக கணிப்பொறி, கணிப்பொறியை, கணிப்பொறியில், கணிப்பொறியால், கணிப்பொறிக்கு போன்ற சொற்கள் தனிச்சொற்களாகவே இடம்பெறும். இவைகளை சொல் வடிவங்கள் என் மொழியியலாளர்கள் கூறுகின்றனர். ஆனால் அகராதியில் இவைகள் தனிச்சொற்களாக இடம்பெறாது. கணிப்பொறி என்ற பெயர்ச்சொல் மட்டுமே அகராதியில் இடம்பெறும். இதுபோன்ற அகராதிச்சொற்களை மொழியியலாளர்கள் Lexemes என்கின்றனர். இது போன்று படி என்ற வினைச்சொல்லுக்கு படித்தான், படித்தார், படித்தது போன்ற எண்ணற்ற சொல் வடிவங்கள் இருப்பினும் படி என்பதே அகராதி சொல்லாகும். கணிப்பொறி வழி சொல்லடைவு தயாரிக்கும் போது சொல் வடிவங்களே இடம்பெறும். இதிலிருந்து அகராதி சொற்களை மனித முயற்சியினாலேயே மேற்கொள்ள வேண்டும். இப்பணியையும் முழுமையாக கணிப்பொறி படுத்த உருபனியல் ஆய்வு மேற்கொள்ள வேண்டும். இவ்வாறாக சொல் வடிவத்திலிருந்து (Word forms) உருபனியல் பகுப்பாய்வு (Morphological Analysis) முலம் அகராதிச் சொற்களை Lexemes – Distionary Words உருவாக்கும் முறையை மொழியியலாளாகள் Lemmatization என்று அழைக்கின்றனர். இம்முறையைக் கணிப்பொறிப்படுத்த முயற்ச்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. இதற்கு இயற்கை மொழி ஆய்வு உதவுகிறது.

முடிவுரை:

கணிப்பொாறி வழி அகராதி தயாரிக்கும் படிநிலைகளில் கணிப்பொறிப்படுத்த இயலும் நிலைகளையும், முறைகளையும் சுருக்கமாகப் பார்த்தோம். இம்முறைகளில் அகராதியின் மொழி ஒரு தடையில்லை. தமிழ் அகராதியோ, ஆங்கில அகராதியோ அல்லது பிற மொழி அகராதியோ மேற்கண்ட முறைகளைப் பின்பற்றி உருவாக்கலாம். அந்தந்த மொழிகளில் அகரவரிசைப்படுத்தப்படும் மென்பொருள்களை வடிவமைக்க வேண்டும்.

கணிப்பொறி வழி மொழி ஆய்வு - தேவைகளும் சிக்கல்களும்

முன்னுரை:

கணிப்பொறியில் ஆங்கிலத்தைத் தவிர பிற மொழிகளையும் கையாளுதல் தொடங்கி உள்ளது. குறிப்பாக கணிப்பொறியை தமிழில் கையாளுதல் கடந்த 10 ஆண்டுகளுக்கு முன் தொடங்கப்பெற்று வளர்ச்சியடைந்து வருகிறது. பொதுவான பயன்பாடுகளுக்கு தமிழைக் கையாளுதல் கணிப்பொறியில் சமுக மேம்பாட்டைப் பயன்படுத்த உதவுகிறது. இது தவிர மொழி ஆய்வு பணிகளுக்காகவும் கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

கணிப்பொறியில் மொழிகள் இருவிதமாக கையாளப்படுகின்றன.

1. அன்றாட பணிகளுக்காக கணிப்பொறியில் மொழியைக் கையாளுதல்
2. மொழி ஆய்வுக்காக கணிப்பொறியைக் கையாளுதல்

முதல் வகை பயன்பாட்டில் கணிப்பொறியில் மொழிகள் கையாள்வது கீழ்கண்ட பணிகளுகட்கு உதவுகின்றன.

1. தமிழ்ச் சொல்லாய்வுகள் உருவாக்கப்படுதல், இதன்மூலம் கணிப்பொறியின் ஒரு உயர்ரக பின்ம தட்டச்சுப் போன்ற தமிழுக்குப் பயன்படுத்துதல்.
2. டி.டி.பி போன்றவைகளுக்கான மென்பொருள்களை தமிழுக்கு உருவாக்கி அச்சுப் பணிகளுக்கு கணிப்பொறியைப் பயன்படுத்துதல்.
3. நிறுவனங்களில் பணியாளர்கள் பட்டியல்கள், ஊதியப்பட்டியல்கள் மற்றும் வரவு செலவு அறிக்கைகளைத் தமிழில் தயாரித்தல்.
4. வணிக நிறுவனங்களில் பொருள் வாங்கும் மற்றும் பொருள் விற்கும் பட்டியல்களை கணிப்பொறி வழி தமிழில் தயாரித்தல்.
5. தொலைவரி, பேஜர் பேன்ற கருவிகளில் செய்தித் தொடர்புக்கு தமிழைக் கையாளுதல் மேற்கண்ட பணிகளுக்காக கணிப்பொறி தமிழைக் கையாளுவதற்காக தொடங்கியுள்ளது. இத்தகைய வசதிகள் பொது மக்களுக்கு பயன்தரத்தக்க வகையில் அமைந்துள்ளன. எனவே இத்தகைய மென்பொருள்களை வடிவமைப்பதற்கு வணிக ரீதியில் செயல்படும் நிறுவனங்கள் முன் வருகின்றன. அரசும் ஆதரவு தருகிறது. இத்தகைய பணிகளுக்கான மென்பொருள்களை வடிவமைப்பதற்கு தற்போது கணிப்பொறியில் உள்ள வசதிகளே போதுமானதாகும். மேலும் சாதாரண மொழி அறிவே மென்பொருள் வடிவமைப்பிற்கு போதுமானதாகும். ஆனால் கணிப்பொறியை மொழி ஆய்விற்கு பயன்படுத்தும்போது மொழியில் புலமையும், செறிவும் தேவைப்படுகின்றன. இத்தகைய பணிகளைக் கணிப்பொறி வல்லுநர்களும், மொழியியலாளர்களும் இலக்கியவாதிகளும் இணைந்து செயல்பட்டு வடிவமைக்கப்பட வேண்டும்.

கணிப்பொறியில் தமிழில் உள்ளீடு செய்வதற்கு வசதிகள் உள்ளன. அதுபோல் வெளியீடு பெறுவதற்கு உயர்ரக அச்சுப் பொறிகள் உள்ளன. ஆனால் மொழி ஆய்வுப் பணிகளை செய்வதற்கு போதுமான மென்பொருள்கள் இல்லை. இதற்கான முயற்சிகள் தொடங்கியுள்ளன.

தமிழ் மொழியாய்விற்கு தேவைப்படுவன.

மொழி ஆய்விற்குப் பின்வரும் மொழிக்கருவிகள் தேவைப்படுகின்றன.

1. அகராதிகள்
2. சொற்களஞ்சியங்கள்
3. சொல் வங்கி (Data Bank)
4. தகவல் தளம் (Data Base)
5. கிளைப்பட வங்கி (Tree Bank)
6. பெருந்தரவு (Corpus)

எந்தவொரு மொழியாய்வு பணிகளுக்கும் மேற்கண்ட கருவிகள் மிக இன்றியமையாததாகும்.

அகராதி தயாரிக்கும் பணியில் உயர்ந்தபட்சமாக சொல்லடைவு வரை கணிப்பொறியில் எளிமையாக தயாரிக்கலாம். சொல்லடைவில் சொற்கள், சொல்வடிவம் என்ற நிலையில் உள்ளன. இதிலிருந்து அகராதிச் சொற்களை தயாரிப்பதற்கு உருபனியல் பகுப்பாய்வு தேவைப்படுகிறது. இப்பகுப்பாய்வினை செய்வதற்கு செயல்முறைகள் கணிதக்கூறுகளாக வடிவமைக்கப்பட வேண்டும். இது எளிய பணி அல்ல. இருப்பினும் இப்பணி மொழி ஆய்வு தொடக்கப்பணியாக அமைகிறது.

தமிழில் அமைந்துள்ள செய்திகளை கணிப்பொறியில் உள்ளீடு செய்வதற்கு விசைப்பலகை பயன்படுத்தப்படுகிறது. தமிழ் இலக்கியங்கள் பிற படைப்புகளை கணிப்பொறியில் தட்டச்சு செய்து உள்ளீடு செய்ய நிறைய பொருட்செலவும், கால விரையமும் ஆகும். எனவே அச்சு வடிவில் அமைந்த தமிழ்ச்செய்திகளையும், கையெழுத்து செய்திகளையும் கணிப்பொறி படித்து சொல் வடிவங்களாக மாற்ற 'OCR'(Optical Character Recognizer) என்ற தொழில் நுட்பம் பயன்படுத்தப்பட்டு ஒலி வருடிக்கான (Scanner) மென்பொருள்கள் உருவாக்கப்பட்டு வருகின்றன. இத்தொழில் நுட்பம் வெற்றி பெறும்பொழுது தமிழ்ச் செய்திகளை உள்ளீடு செய்வது மிக எளிய பணியாக அமையும்.

கணிப்பொறிக்கு தரவேண்டிய கட்டளைகள் மனித மொழியில் தருவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. ஒரு மனிதன் மற்றொருவரிடம் பேசுவது போல் உரையாடுவது போல் கணிப்பொறியுடன் உரையாடுவதற்கு மென்பொருள்கள் உருவாக்கப்பட்டு வருகின்றன. இப்பணிகளுக்கு எல்லாம் எண்ணற்ற பகுப்பாய்வுகளும் ஒத்து சேரும் மென்பொருள்கள் தேவைப்படுகின்றன. இப்பணிகள் எல்லாம் மொழியியலாளர்கள் கணிப்பொறியாளர்கள் கூட்டு முயற்சியால் நடைபெறுகிறது. தானியங்கி மொழிபெயர்ப்பிற்கு ஒரு மொழியின் இலக்கணக்கூறுகள் அனைத்தையும் கணிப்பொறிக்குத் தர வேண்டி உள்ளது. அதாவது கணிப்பொறிக்கு மூல மொழி (Source Language) மற்றும் இலக்கு மொழி (Target Language) இரண்டிலும்

தேவையான அறிவினை ஊட்ட வேண்டும். இப்பணிகளும் கணிப்பொறி தயாரிப்பாளர்களின் கூட்டு முயற்சியிலேயே நடைபெறுகின்றன.

மொழியாய்வில் மொழியியலாளர்களுக்கு தேவைப்படுவன.

1. கணிப்பொறியைப் பற்றிய அடிப்படை அறிவு
2. மொழியின் இலக்கணம் மற்றும் அதனை கடினப்படுத்துவதில் திறன்.
3. ஒரு திட்டத்தில் எப்பணிகளை கணிப்பொறி வழி செய்ய இயலும் என்பதற்கான திட்டம்.
4. ஒன்றிணைந்து பணிபுரியும் கணிப்பொறியாளரிடம் மொழி சிக்கல் பற்றியும் தீர்வு பற்றியும் எடுத்துரைக்கும் திறன்.
5. சிக்கலுக்கான தீர்வினை படிப்படியாக பல்வேறு நிலைகளில் செயல்படுத்துவதற்கான திறன்.
6. கணிப்பொறிபடுத்தும் போது தோன்றும் சிக்கல்களை சக மொழியாளர்களிடம் காண்பித்து அதற்கு தீர்வு காணும் திறன்.

கணிப்பொறி மொழியாய்வில் கணிப்பொறியாளர்க்குத் தேவைப்படுவன.

1. மொழியிலும் மொழி இலக்கணத்திலும் அடிப்படை அறிவு பெற்றிருத்தல் வேண்டும்.
2. மொழியாய்வு பணியை பல்வேறு சிறு சிறு பணிகளாகப் பிரித்து கணிப்பொறி வழி செயல்படுத்துவதற்கான மென்பொருள்களை உருவாக்கும் திறன்.
3. மொழி ஆய்வு பணிகளுக்கு உதவும் கணிப்பொறி வழியமைப்புகளில் தேர்ச்சி பெற்றிருத்தல்.
4. மொழியாளர்களுக்குத் தேவைப்படும் மாதிரி வழியமைப்புகளை உருவாக்க மொழியாளர்களுக்கு செயல் விளக்கம் தந்து அவர்களின் ஒத்துழைப்பைப் பெறுதல்.
5. மொழியாளர் என்னும் பிரச்சனைகளுக்கான தீர்வினை உடனுக்குடன் எதிர்பாராமல் இருத்தல் மொழியாளர்களுக்கு கணிதப்படுத்துவதில் பயிற்சியின்மையால் உடனுக்குடன் தீர்வினை தர இயலாது. இந்நிலையில் அவர்களுள் ஒத்துழைத்து தீர்வு பெற முயற்சித்தல்.
6. மொழி ஆய்வு பணிகளுக்கென அவ்வப்பொழுது வெளிவரும் மென்பொருள் மற்றும் கணிப்பொறி வசதிகளைத் தெரிந்துகொள்ளுதல்.

மேற்கூறியவாறு மொழியியல் ஆய்வு பணிகளுக்கு மொழியாளர் மற்றும் கணித பொறியாளர்க்கு ஒத்துழைப்பு.

நிறைவு:

கணிப்பொறி வழி மொழியாய்வு பணியினை மேற்கொள்வதற்குத் தேவைப்படும் மென்பொருள்களை உருவாக்க கணிப்பொறியாளர்கள், மொழியியலாளர்கள், இலக்கியவாதிகள் ஆகியோரின் கூட்டு முயற்சிகள் தேவையாகும்.

கணிப்பொறி வழி மொழிபெயர்ப்பு (Machine Translation

பாரதத்தில் பன்மொழி பேசும் மக்கள் உள்ளன. மொழி, இனம், பண்பாடு மற்றும் கலாச்சாரத்தால் பல்வேறு வகையாக உள்ளனர். ஒரு மொழி பேசும் மக்களின் எண்ணங்கள், கருத்துக்கள் பிற மொழி பேசும் மக்களிடையே சென்றடைவது அவசியமாகும். அது போல் ஒரு மொழி பேசும் மக்களின் பண்பாடு, கலாச்சாரம், இலக்கியம் போன்றவைகள் பிற மொழி பேசும் மக்களிடையே சென்றடைய வேண்டும். அப்பொழுதுதான் மக்களிடையே ஒற்றுமை உணர்ச்சியும், தேசிய ஒருமைப்பாடும் நிகழும். எனவே ஒரு மொழியில் உள்ள செய்திகள் பிற மொழிகளில் மாற்றம் செய்யப்படுவது மிகவும் அவசியமாகும்.

இது போல் உலக நாடுகளில் ஒரு நாடு பிற நாடுகளில் இருந்து தனித்திருக்க வேண்டும். தொழில் நுட்ப மேம்பாடு, பொருளாதார மேம்பாடு, சமூக மேம்பாடு போன்றவைகளுக்கு ஒன்றோடு ஒன்று தொடர்புக் கொண்டு கூட்டு முயற்சியில் மனித வாழ்வியல் மேம்பாட்டினை கருத்தில் கொள்ளவும் வேண்டும். ஒரு நாட்டின் கலாச்சாரம், பண்பாடு போன்றவைகள் பிற நாடுகளுக்கும் பரவச் செய்ய வேண்டும். ஒரு நாட்டில் இலக்கியச் செல்வங்கள் பிற நாடுகளைச் சென்றடைய வேண்டும். அப்பொழுதுதான் ஒரு நாட்டின் தொன்மை, தனித்துவம், வளம் போன்றவைகள் பிற நாடுகளுக்கு தெரிய வரும். எனவே ஒரு நாட்டில் உருவாக்கப்படும் படைப்புகள் பிற நாடுகளுக்கு எளிதில் சென்றடைய வேண்டுமெனில் மொழிபெயர்ப்புப் பணி அவசியமாகும்.

நாம் தகவல் யுகத்தில் வாழ்ந்து வருகிறோம். செய்திகள் மலையெனக் குவிகின்றன. இச்செய்திகளை உடனுக்குடன் மொழிபெயர்க்க வேண்டியது அவசியமாகிறது. வாழ்வில் எல்லா நிலைகளிலும் பயன்தரக்கூடிய கணிப்பொறி மொழிபெயர்ப்புகளை பயன்படுத்த முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன.

கணிப்பொறி வழி மொழிபெயர்ப்பு தேவை

ஒரு மொழியியல் உள்ள செய்திகளை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழிபெயர்ப்பு எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியில் இயல்பு தன்மை மாறாமல் இருத்தல் வேண்டும். செய்தி எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Language) என்கிறோம். செய்திகள் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அதை அந்த மொழியை இலக்கு மொழி (Target

Language) என்கிறோம். இத்தகைய மொழிபெயர்ப்புகளை மனித முயற்சியில் மேற்கொள்வது என்பது மிகக் கடினமான பணியாகும். எனவே மொழிபெயர்ப்புகளை கணிப்பொறி வழி மேற்கொள்வது அவசியமாகும். கணிப்பொறி வழி மொழிபெயர்ப்பு மேற்கொள்வதற்கான தேவைகள் மற்றும் அவசியங்கள் கீழே பட்டியலிடப்பட்டுள்ளன.

1. உலகளவில் செய்திகள் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழிபெயர்த்தல் அவசியமாகும்.
2. கணிப்பொறி வழி மொழிபெயர்த்தல் மனித உழைப்பைக் குறைக்கிறது.
3. குறைந்த நேரத்தில் நிறைய மொழிபெயர்ப்பு பணிகளை செய்ய இயல்கிறது. இதனால் காலவிரயம் தவிர்க்கப்படுகிறது.
4. தற்போதைய மொழிபெயாப்பு பணிகளுக்கு ஏற்ற வகையில் மொழிபெயர்ப்பாளர்கள் இல்லை.
5. மனித முயற்சியில் மொழிபெயர்கின்ற பொழுது மொழிபெயர்ப்பாளர் மூலமொழி மற்றும் இலக்கு மொழிகளில் புலமை வாய்ந்தவராக இருத்தல் வேண்டும். ஆனால் கணிப்பொறிக்கு இந்தக் கட்டுப்பாடு இல்லை. தேவையான மொழியில் புலமை உள்ளவர்களிடம் கருத்துக்கள் பெற்று அதை வழியமைப்பு மொழிகளில் எழுதி மொழிபெயர்ப்பு செய்யலாம்.

கணிப்பொறி மொழிபெயர்ப்பின் வகைகள்:

பொதுவாக கணிப்பொறி வழி மேற்கொள்ளப்படும் மொழிபெயர்ப்பு பின்வரும் மூன்று வகைகளாகப் பிரிக்கப்படுகிறது.

1. கணிப்பொறி உதவியுன் மனித மொழிபெயர்ப்பு (CAHT – Computer Aided Human Translation)
2. மனித உதவியுடன் கூடிய கணிப்பொறி மொழிபெயர்ப்பு (HAMT- Human Aided Machine Translation)
3. முற்றிலும் தானியங்கி மொழிபெயர்ப்பு (AMT – Automatoc Machine Translation)

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனித மொழிபெயர்ப்பில் மனித மொழிபெயர்ப்பாளருக்கு முக்கியத்துவம் தரப்படுகிறது. அவர்கள் மொழிபெயர்ப்பு பணியில் ஈடுபடும் போது ஒரு மொழியாக உள்ள சொற்களுக்கு நிகரான பிற மொழிச்சொற்களைத் தெரிந்துகொள்ளவும் அச்சொற்களின் பொருள் பயன்பாடு ஆகியவற்றை எளிமையாக தெரிந்து கொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு,

தொடர் அமைப்பு போன்றவைகளுக்கு நிகரான பிற மொழி அமைப்புகளை விரைவாக தயார் நிலையில் தெரிந்து கொள்வதற்கு கணிப்பொறி உதவுகிறது. இவ்வகை மொழிபெயர்ப்புகளில் மொழிபெயர்ப்பாளர்கள் முதல் நிலையிலும் அவர்களுக்கு உதவக்கூடிய வகையில் கணிப்பொறி உதவுகிறது.

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிபெயர்ப்பு பணியை செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ள தகவலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிபெயர்ப்பு வகைகளில் மனித உதவியுடன் நடைபெறும் இயந்திரமொழிபெயர்ப்பே நடைமுறையில் உள்ளது.

முற்றிலும் இயந்திர மொழிபெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல. இதற்கு எண்ணற்ற மொழியியல் கோட்பாடுகளும், மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கபட வேண்டி உள்ளது. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பில் முற்றிலும் தானியங்கப்படுத்தும் முயற்சி இலக்காக கொள்ளப்பட்டுள்ளது.

கணிப்பொறி வழி மொழிபெயர்ப்பின் வரலாறு

கணிப்பொறி வழி மொழிபெயர்பினை தானியங்கி மொழிபெயர்ப்பு Automatic Translation என்றும், இயந்திரமொழிபெயர்ப்பு Machine Translation என்றும் கணிப்பொறி உதவியுடன் நடைபெறும் மொழிபெயர்ப்பு என்றும் கொள்ளலாம். கணிப்பொறியின் வேகம், நினைவகத்திறன், பதிவு செய்துள்ள செய்திகளை தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளப் பெறக்கூடிய வசதி (Retriev) போன்றன கணிப்பொறி வழி மொழிபெயர்ப்பு பணியை மேற்கொள்ள உதவின.

கணிப்பொறி வழி மொழிபெயர்ப்பு பணி 1950 ஆம் ஆண்டு வாரன்வீவர் (Warrenweiver) என்ற அறிஞரால் தொடங்கப்பட்டது. இதைத் தொடர்ந்து பு. பார்ஷில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்பு பணி ஆரம்பித்தார்.

1954 ஆம் ஆண்டு ஜார்ஜ்டவுன் பல்கலைக்கழகம் கணிப்பொறி வழி மொழிபெயர்ப்பு பணியைத் தொடங்கியது. இதைத் தொடர்ந்து CUTA (Chinese University Translation) என்ற மென்பொருள் உருவாக்கப்பட்டது. இதைக் கொண்ட சீனமொழியில் உள்ள செய்தித்தாள்கள் ஆங்கில மொழிக்கு மொழி மாற்றம் செய்யப்பட்டன. 1947 ஆம் ஆண்டு Systran என்ற மென்பொருள் உருவாக்கப்பட்டது. Systran என்பதன் விரிவு System

Translation என்பதாகும். இம்மொழிபெயர்ப்பு மென்பொருள் ரஷ்ய மொழியிலிருந்து பிரெஞ்சு மொழிக்கு மொழி மாற்றம் செய்ய பயன்பட்டது. இதற்கான திட்டத்தை P.Toma என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன்முதலில் வெளியிடப்பட்ட கணிப்பொறி வழி மொழிபெயர்ப்பிற்கான மென்பொருள் இதுவே ஆகும்.

1975 ஆம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென்பொருள் ரஷ்ய மொழியில் அமைந்துள்ள செய்திகளை பிரெஞ்சு மொழிக்கு மாற்றம் செய்யப்படுவது. இதே காலக்கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு மொழிபெயர்க்க வடிவமைக்கப்பட்டது.

இதுபோன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனை கட்ட முயற்சியாகவும் வணிக நோக்கத்திலும் கணிப்பொறி வழி மொழியபெயர்ப்பதற்கான மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைபெற்று வருகின்றன.

1985 ஆம் ஆண்டு தமிழ்ப்பல்கலைக்கழகத்தில் TUMTS – Tamil Universoty Machine Translation System என்ற திட்டப் பணி தொடங்கப்பெற்றது. ரஷ்ய மொழியில் உள்ள செய்திகளை தமிழில் மொழிபெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழக கணிப்பொறி அறிவியல் துறை, மொழியியல்துறை மற்றும் மொழிபெயர்ப்புத்துறை ஆகியவைகளின் கூட்டு முயற்சியால் விளைந்ததாகும்.

இதற்கு பின்னால் கான்பூரில் உள்ள இந்திய தொழில் நுட்ப நிறுவனம் (Indian Institute of technology – IIT) ஹிந்தியிலிருந்து ஆங்கிலம், ஆங்கிலத்திலிருந்து ஹிந்தியை மொழிபெயர்க்க வடிவமைக்கப்பட்டுள்ளது.

தற்பொழுது சென்னையிலுள்ள AUKRC – Anna University Kalanasundaram Research Center என்ற ஆய்வு நிறுவனம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்கத் திட்டம் ஒன்றினை மேற்கொண்டு செயல்படுத்தி வருகிறது.

கணிப்பொறி வழி மொழிபெயர்ப்பு - எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிகளுக்கு மொழி மாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்தகைய கணிப்பொறி வழி மொழிபெயர்ப்புப் பணிகளில் பின்வரும் இரு எல்லைகள் நிர்ணயிக்கப்படுகின்றன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் அமைப்புகளின் தொகுப்பு ஆகும். ஒரு வட்டாரத்தில் பயன்படுத்தக்கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்துவது என்பதற்கோ அதே பொருளில் பயன்படுத்துவது

என்பதற்கோ உறுதி இல்லை. தமிழ் போன்ற மொழிகளில் மொழி இலக்கிய மொழி, பேச்சு மொழி என இரு வகையாகப் பிரிக்கப்படுகிறது. பேச்சு மொழியை கணிப்பொறியாக்கம் செய்வது என்பது சுலபமான பணி அன்று எனவே மொழிபெயர்ப்பு பணிகளில் இலக்கியம் மட்டுமே எடுத்துக்கொள்ளப்படுகின்றன.

2. செய்யுள், கவிதை போன்றவைகளில் பொருளுக்குத் தரும் முக்கியத்துவத்தைவிட கருத்தினைப் புலப்படுத்தும் வசதி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகின்றன. இவைகள் மனித உளவியலை பிரதிபலிப்பன. இவ்வுளவியல் கூறுகளை கணிப்பொறி வழி வெளிப்படுத்துவது என்பது இயலாத ஒன்றாகும். இதுபோல் இலக்கியங்களில் பொருளைவிட உளவியல் கூறுகளுக்கு முக்கியத்துவங்கள் தரப்படுகின்றன. எனவே செய்யுள், கவிதை, பாடல்கள் மற்றும் இலக்கிய செய்திகள் ஆகியவைகளைக் கணிப்பொறி வழி மொழிபெயர்ப்பிற்கு தொடக்க நிலையில் எடுத்துக்கொள்ள வேண்டும். அறிவியல், தொழில்நுட்பம் சார்ந்த படைப்புகளில் பொருளுக்கே முக்கியத்துவம் தரப்படுகிறது. இங்கு கோபம், வீரம், தாபம், எழுச்சி போன்ற உணர்வுகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழிபெயர்ப்பு பணிகளுக்கு பொதுவாக அறிவியல் மற்றும் தொழில்நுட்பத்தைச் சார்ந்த படைப்புகளே எடுத்துக்கொள்ளப்படுகின்றன.

மேற்கூறிய இரண்டும் கணிப்பொறி மொழிபெயர்ப்பிற்கு எல்லைகளாகக் கொள்ளப்படுகின்றன.

உருஷ்யன் - தமிழ் மொழிபெயர்ப்பு ஒரு கண்ணோட்டம் அல்லது முன்னேற்பு பணிகள்:

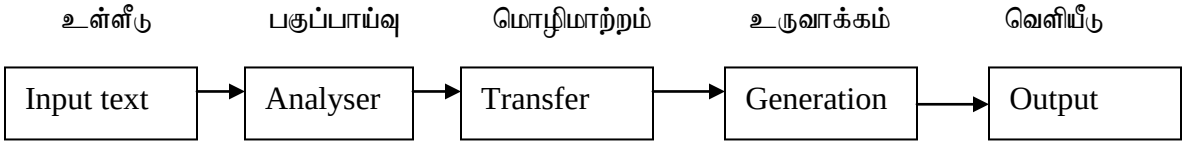
இரு மொழிகளுக்கு இடையே உள்ள சொல் வளம் (Vacabulary Grammatr) இலக்கணம் கருத்தாடல் (Concept Structure) உள்ள ஒற்றுமை வேற்றுமைகளை பற்றி ஆராயப்படும் பொழுது ஒற்றுமை பற்றி ஆராயும் பிரிவுக்கு (Typology) என்று பெயர். இரு மொழிகளுக்கு இடையே உள்ள வேற்றுமைகளை மட்டும் அறிந்து கொள்வதற்கு Contrast Knoweldge என்று பெயர். உருசியன் - தமிழ் மொழிபெயர்ப்பு பணியில் இவ்விரு மொழிகளுக்கு இடையே உள்ள சொல்வளம் சொற்களுக்கு இடையே உள்ள உறவுகள் தொடரியல் மற்றும் வாக்கியங்களுக்கு இடையே உள்ள உறவுகள் கண்டறியப்பட்டு வரைமுறைப்படுதப்பட்டன. இம்மொழிபெயர்ப்பிற்கு பின்வரும் இரு குறிக்கோள்கள் அடிப்படையாகக் கொள்ளப்பட்டன.

1. எழுத்து அல்லது இலக்கிய உரைகளை மட்டும் மொழிபெயர்ப்பிற்கு எடுத்துக்கொள்வதால் பேச்சு மொழி வழக்கினை ஒதுக்குவது.

2. இலக்கிய படைப்புகளைத் தவிர்த்து அறிவியல் மற்றும் தொழில் நுட்ப உரைகளை மட்டும் எடுத்துக்கொள்வது.

முன்னேற்புப் பணிகள்

1. இம்மொழிபெயர்ப்பிற்கு ருஷ்ய மொழியை மூலமொழியாகவும் (Source Language) தமிழை இலக்கு மொழியாகவும் (Target Language) எடுத்துக்கொள்ளப்பட்டன.
2. மூலமொழி அதாவது ருஷ்யன் மொழியில் அமைந்துள்ள செய்திகள் ஒலிப்பெயர்ப்பு (Transliteration) செய்து கணிப்பொறிக்கு உள்ளீடு செய்வது.
3. உருசிய மொழிக்கும் தமிழுக்கும் ஒரு அகராதி உருவாக்கப்பட்டது.
4. உருசிய மொழியில் அமைநதுள்ள சொற்கள், சொற்றொடர்கள் மற்றும் வாக்கிய அமைப்புகளின் சில முன்திருத்தம் (pre processing) செய்தல்.
5. மூல மொழியில் உள்ள சொல் மற்றும் வாக்கிய அமைப்பிற்கு நிகரான தமிழ்ச்சொல் வாக்கியம் ஆகியவற்றை நிர்ணயம் செய்ய விதிமுறைகளை உருவாக்குவது. இத்தகைய விதிமுறைகளும், வழிமுறைகளும் மாற்றம் (Transfer) என்ற பெயரில் பதிவு செய்தல்.
6. கணிப்பொறி வழி மொழிபெயர்ப்பு செய்த பின்னர் தேவைப்படின் தமிழ் மொழிக்கேற்ப வாக்கிய அமைப்புகளை திருத்தி அமைப்பது (Post Processing)



உருசியன் - தமிழ் மொழிபெயர்ப்பு படி நிலைகள்

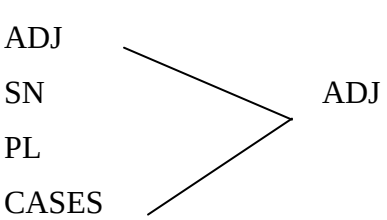
உருசியன் தமிழ் மொழிபெயர்ப்பிற்கான படிநிலைகளின் விளக்கம் கீழே தரப்பட்டுள்ளன.

மூலமொழி உருசியனை L1 என்றும் தமிழ் மொழியை L2 என்றும் குறிப்பிடுகின்றன. முதற்கட்டமாக மூலமொழியில் உள்ள ஒரு செய்தி ஒலிபெயர்ப்பு செய்யப்பட்டு கணிப்பொறிக்கு உள்ளீடு செய்கின்றன.

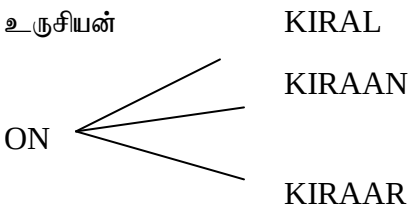
இரண்டாம் கட்ட நிலையில் மூல மொழியில் உள்ள செய்தி பகுப்பாய்வு செய்யப்படுகிறது. இந்நிலையில் சொற்களின் இலக்கண கூறுகள் கண்டறியப்படுகின்றன. இதற்கு ருஷ்யன் - தமிழ் அகராதிகள் உதவுகிறது. இப்பகுப்பாய்வு மூலம் ருஷ்ய மொழியில் உள்ள செய்திகள் NP, VP, NN, ADF என பாகுபடுத்தப்படுகின்றன.

மூன்றாம் கட்ட நிலையில் பகுப்பாய்வு செய்யப்பட்டுள்ள மூலமொழியின் வார்த்தையில் உள்ள சொற்களின் படிவம், வாக்கிய அமைப்பு போன்றவைகளுக்கு இலக்கு மொழியில் நிகராக

உள்ள சொற்களின் வடிவம், வாக்கிய அமைப்பு ஆகியன Transfer என்ற அமைப்பின் மூலம் கண்டறியப்பட்டு மொழி மாற்றம் செய்யப்படுகின்றன. இந்நிலையில் இவ்விரு மொழிகளுகட்கும் இடையே உள்ள Contrast Knoweldge பயன்படுத்தப்படுகின்றன. கருத்தாடலைப் பொருத்தவரை இவ்விரு மொழிகளுக்கு இடையே One – to – Many உறவு உள்ளது. பெயர்வரிச் சொல்லை பொறுத்தவரையில் இவ்விரு மொழிகளுக்கு இடையே Many – to – One உறவு உள்ளது. இவ்வுறவினை கீழ்கண்ட படம் விளக்குகிறது.



மேலும் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமை உருபுகளில் ONE – TO – MANY உறவு உள்ளது.



இதுபோன்ற இவ்விரு மொழிகளுக்கு இடையே உள்ள வேறுபாடுகள் விதிமுறைகளாகவும், வழிமுறைகளாகவும் (Transfer) என்ற அமைப்பின் கீழ் உள்ளன. பகுப்பாய்வு செய்யப்பட்ட மூலமொழி செய்திகள் (Transfer) அமைப்பில் உள்ளீடாக செல்கிறது. இந்நிலையில் மூலமொழி அமைப்பிற்கு நிகரான இலக்கு மொழி அமைப்பிற்கு செய்தி மொழிமாற்றம் செய்யப்படுகிறது.

நான்காவது கட்ட நிலையில் மொழி மாற்றம் செய்யப்பட்டுள்ள இலக்கு மொழி கட்டமைப்புகள் இலக்கு மொழியின் தன்மைக்கேற்ப அமைவதில்லை. இவ்வாறு கிடைக்கப்பெற்ற இடைநிலை மொழிமாற்ற அமைப்புகளை இலக்கு மொழியில் தன்மைக்கேற்ப உருவாக்கித் தருவது தான் (Genaration) என்ற பகுதியின் பணியாகும். இந்நிலையில் மொழிமாற்றம் செய்யப்பட்ட செய்திகள் உள்ளன. இச்செய்தியை மூலமொழியின் வரிவடிவத்திலே அல்லது ஒலிபெயாப்பு வடிவத்திலே தரும் பணியை Out put என்ற அமைப்பு தருகிறது. மொழிபெயர்ப்பு செய்யப்பட்ட செய்தியை காட்சித் திரையில் காணலாம். தேவைப்படின்

அச்சுப்பொறி வழி அச்சிட்டுக்கொள்ளலாம். உருசியன் - தமிழ் மொழிபெயர்ப்பை பொருத்தவரையில் வெளியீட்டை தமிழ் எழுத்துக்களில் காண்பதற்கும், அச்சிட்டுக் கொள்வதற்கும் வசதிகள் உள்ளன. அதுபோல் ஒலிபெயர்ப்பு முறையிலும் விடையை பெறலாம்.

கணிப்பொறி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழிபெயர்ப்பு பணியை கணிப்பொறி வழி மேற்கொள்கின்ற பொழுது கணிப்பொறியின் சிறப்புத் தன்மைகளான அதிவேகம், மிகுந்த நினைவக கொள்திறன், தகவல் தளம் உருவாக்கும் வசதிகள் போன்றவை திறமையாக பயன்படுத்தப்படுகின்றன. ஒரு மொழியில் இருந்து பிற மொழிக்கு மொழிபெயாப்பு செய்யப்படும் பொழுது பன்மொழித்திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வடிவமைக்கப்படுகின்றன. எனவே இருமொழிபெயர்ப்பு பணிக்குக்கூட ஒருவருக்கு இருமொழிகளிலும் திறமை இருக்க வேண்டும் என்ற அவசியம் இல்லை. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பு பணிகளில் பல சிக்கல்கள் உள்ளன. ஒரு மொழியின் வளம், தன்மை ஆகியவற்றை கணிப்பொறிக்கு ஏற்றார்போல் கணிதப்படுத்துவது என்பது எளிமையான பணி அல்ல. எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத் திட்டங்கள் சுமார் ஐம்பது ஆண்டுகளுக்கு முன்னர் தொடங்கப்பெற்றிருப்பினும் இன்னும் முழுமைப்பெற வில்லை. மேலும் கணிப்பொறி அறிவியலின் தொழில்நுட்ப வளர்ச்சி கணிப்பொறியில் மொழிபெயர்ப்பு பணிக்கு மிகவும் பயனுள்ளதாக அமைகின்றன. இத்தகைய பணிகளின் குறை நிறைகளை கீழே காண்போம்.

1. மொழிகளுக்கு இடையே உள்ள சொல்வளம், இலக்கணம், கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவ்வகைகளில் மொழிகளுக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிபெயர்ப்பு பணியின் சிக்கல்கள் குறையும்.
2. உருபனியல் அமைப்பைப் பொருத்தவரையில் மொழிகளின் குணாதிசியங்கள் இருவேறு திசைகளில் மாறுகின்றன. 1. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதை பொருத்து. 2. ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருக்குமேயானால் அவைகளை எவ்வளவு துல்லியமாக பிரிக்க இயலும் என்பதைப் பொருத்து ஒரு சொல்லில் ஒரே உருபன் மட்டும் இருந்தால் ISOLATING மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களின் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை Polysyntietic என்கிறோம். சிபேரியன் மொழி Polysyntietic

மொழிக்கு சிறந்த உதாரணமாகும். வியட்நாமில் மற்றும் கேன்டுனீஸ் ஆகிய மொழிகளில் Isolating மொழிக்கு சிறந்த உதாரணங்களாகும். எனவே இத்தகைய மொழிக்குள் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் உருவாக்குகின்ற பொழுது மொழிகளின் உருபனியல் பற்றிய விரிவான தகவல் தளம் அவசியமாகும்.

3. மொழிகளின் தொடரியல் அமைப்பைப் பொருத்து (Syntax Structure) மொழிகள் முன்று வகைகளாக பிரிக்கப்படுகின்றன.

1. SVO அமைப்பு

ஜெர்மன், பிரெஞ்சு, ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

ஹிந்தி, ஜப்பான் போன்ற மொழிகள் இந்த அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரிஸ், அரேபி, ஹிப்ரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளனவாகும்.

SVO அமைப்பில் உள்ள மொழிகளில் Perapciti பயன்படுகிறது. இதற்கு நேர் எதிராக SOV மொழியில் Post Possication பயன்படுகிறது.
எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத் திட்டங்களில் மொழிக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள் உறவுகள் போன்றவற்றை ஆராய வேண்டும்.

4. இராமனின் வீடு என்ற தொடரில் வீடு என்பது தலைச்சொல் (Headword) ஆகும். இராமன் என்பது சார்புச் சொல்லாகும். தலைச்சொல்லுக்கும் சார்புச்சொல்லுக்கும் இடையில் உள்ள உறவை 'இன்' என்ற ஒட்டு (Affix) விளக்குகிறது. இதுபோன்று தலைச்சொல்லுக்கும் சார்புச்சொல்லுக்கும் உள்ள உறவை ஒட்டு நிர்ணயிக்கும் முறையைப் பொருத்து மொழி (Head) – Marking மொழி Dependent Marking என இரு வகையாகப் பிரிக்கப்படுகிறது. Head Marking மொழியில் ஒட்டு தலைச்சொல்லுடனும் Dependent Marking மொழியில் ஒட்டு சார்பு சொல்லுக்கும் தரப்படுகின்றன. ஆங்கில மொழி Dependent Marking வகையைச் சார்ந்ததாகும். ஹங்கேரியன் மொழி Head Marmking வகையைச் சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு இதை விளக்கும்.

English: Mans house

Huagarian : The man house his

H > Head word

தலைச்சொல்

The man Dependent word

சார்புச் சொல்

A – Affix – ஒட்டு

எனவே கணிப்பொறி வழி மொழிபெயர்ப்புப் பணியை மேற்கொள்கின்றபோது தலைச்சொல் சார்புச்சொற்களுக்கு இடையே உள்ள உறவு முறைகள் எவ்வாறு பிரிக்கப்படுகிறது என்பது பற்றி எல்லாம் ஆராய வேண்டும்.

5. மூலமொழியில் உள்ள ஒரு சொல்லின் இலக்கண கூறு (Part of Speech) அச்சொல்லுக்கு நிகராக இலக்கு மொழியில் உள்ள சொல்லின் இலக்கணக் கூறு ஒன்றாக இருக்க வேண்டிய அவசியமில்லை. வேறுவகையாகக் கூறின் ஒரு மொழியில் உள்ள வினைச்சொல் பிற மொழிக்கு மொழிபெயர்ப்பு செய்கின்ற பொழுது மொழிபெயர்ப்புச் சொல் வினைச்சொல்லாகத்தான் இருக்க வேண்டும் என்ற அவசியமில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்புத் திட்டங்களில் உள்ள சிக்கல்கள் ஆகும்.

6. மொழிகளின் ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கான தளம் விளக்கமானதாகவே சுருக்கமானதாகவே இருக்கலாம். உதாரணமாக சகோதரன் என்ற உறவைக் குறிக்க Brother என்ற சொல் உள்ளது. மூத்த சகோதரனை Elder Brother என்றும் இளைய சகோதரனை Younger Brother என்றும் சொல்கிறோம். எனவே வயதில் மூத்தவர், இளையவர் என்பதைக் குறிக்க Elder, younger என்ற பதங்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் அண்ணன் என்றும் தம்பி என்றும் தனித்தனி சொற்கள் உள்ளன. இங்கு சகோதரன் என்ற உணர்வு மேலும் வகைப்படுத்தப்பட்டு முதியவர், இளையவர் என்ற வகையில் அண்ணன், தம்பி என்ற சொற்கள் உள்ளன. ஆனால் ஆங்கிலத்தில் இவ்வாறான வகை என்பது இல்லை. உறவை மேம்படுத்த Elder, younger என்ற சொற்களைப் பயன்படுத்த நேர்கிறது.

7. மூலமொழியில் உள்ள செய்திகளின் கலாச்சார பின்னணி மொழிபெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்கு மொழிக்கும் பொருத்தமானதா அமைதல் வேண்டும். அல்லது இலக்கு மொழியின் கலாச்சார பண்புகளுக்கு ஏற்ப மாற்றி அமைத்துத்தரதக்க வகையில் இருத்தல் வேண்டும். இரு மொழிகளுக்கு சூழல்களில் சிக்கலைப் புரிந்து கொண்டு அதற்கேற்ற தீர்வினை உரிய

முறையில் வழங்கப்பட வேண்டும். எனவே இயந்திர மொழிபெயர்ப்பு திட்டத்தை இயந்திர – மனித உரையாடலுக்கு ஏற்றாற் போல் வடிவமைக்க வேண்டியது அவசியமாகும். பெரும்பாலும் படைப்பாற்றல் திறன் செய்திகளில் வெளிப்படுத்தும் பொழுது கணிப்பொறிக்கு கூடுதல் செய்திகள் தேவைப்படுகின்றன.

8. பெரும்பாலான கணிப்பொறி வழி மொழிபெயர்ப்புகளின் உள்ளீடு செய்திகள் முன்திருத்தம் செய்யப்படுகின்றன. இத்தகைய முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும் இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. இதை இரு வகையில் செய்யலாம். அ. மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழி பெயர்ப்பது பின்னர் மொழிபெயர்ப்பைப் படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகிறதா என்பதை ஆராய்வது அப்படி இல்லையென்றால் மூலமொழி செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிபெயர்ப்பு செய்வது சரியான கருத்து வெளிவருகின்ற நிலையில் இத்தகைய ஆய்வை மீளச்செய்வது. ஆ. மொழிபெயர்ப்புக்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் கொடுத்து தேவையான திருத்தங்களை செம்மையாக மேற்கொள்வது இத்தகைய நடைமுறை மிகச் சிறந்ததாக கருதப்படுகிறது. இம்முறை கணிப்பொறியின் நேரத்தையும் மிச்சப்படுத்துகிறது. மொழிபெயர்ப்பின் வேகத்தை அதிகமாக்குகிறது. மேலும் ஒரு மொழியில் இருந்து பல மொழிகளுக்கு மொழிபெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப்பணி திறமையானதாக அமைகிறது.

9. மொழிபெயர்ப்பு பணிகளுக்காக உருவாக்கப்படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லது பல கூடுதல் செய்திகளை கொண்டதாக இருத்தல் வேண்டும். பிற வகை அகராதிகளை விட துறை சார்ந்த அகராதிகள் மொழிபெயர்ப்புப் பணியை எளிமைப்படுத்துகிறது. மூலமொழியில் உள்ள செய்தியைப் பொருத்து அது தொடர்புடைய அகராதிகளை பயன்படுத்திக் கொள்ளலாம்.

10. மொழிபெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத்தொடர்கள் மேலும் அடிக்கடி பயன்படுத்தக்கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்பட வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்பு பணி என்பது ஒட்டுமொத்தமாக செய்யக்கூடியது அல்ல. ஆய்வு அடிப்படையில் செய்யக்கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கபடபட வேண்டும். அதை செயல்படுத்தி செயல்திறனை அறிய வேண்டும். பணியின் தன்மையை மதிப்பிட வேண்டும். இத்தகைய எதிரொலிகளுக்கு ஏற்றார் போல் தேவைப்படின் மென்பொருளை கூடுதல் வசதிகளுடன்

மேம்படுத்த வேண்டும். இம்முறைகளை மனதிற்கு நிறைவான செயல்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்ப செய்ய வேண்டும்.

11. கணிப்பொறி வழி மொழிபெயர்ப்பு கடுமையான பணியாகும். இதற்கு மனித மொழிபெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (Models) வழி முறைகளும் (Process) அவசியமாகும். கணிப்பொறி மொழிபெயர்ப்பு ஆக்கபூர்வமான பணியாகும் இங்கு அறிவயல் மட்டுமல்லாது மொழிபெயர்ப்பும் தேவையான கலைத்திறனும் அவசியமாகும்.
12. தற்பொழுது உலகளவில் உருவாக்கப்பட்டுள்ள மொழிபெயர்ப்பு பணிகள் மனதிற்கு நிறைவு தரக்கூடிய வகையில் உள்ளன. இவைகள் மூலம் கிடைக்கப் பெறுகின்ற மொழிகள் எல்லாம் தோராயமான மொழிபெயர்ப்பிற்கு போதுமானது. இவைகளை செம்மைப்படுத்தி முற்றிலும் சிறந்ததாக அமைப்பிற்கு முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. எனவே தற்பொழுது உருவாக்கப்பட்டுள்ள கணிப்பொறி வழி மொழிபெயர்ப்பு பணிகள் வரைவு (Ruft Draft) மொழிபெயர்ப்பிற்கு போதுமானதாகும்.
13. மனிதனின் படைப்பாற்றல், கற்பனைத்திறன், உளவியல் பாங்கு போன்றன செய்திகளில பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணிகளில் இத்தகைய கூறுகளை சேர்க்க வேண்டியுள்ளது. இதற்கு மனித நுண்ணறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன. இத்திட்டங்களில் வெற்றி பெறுகின்ற பொழுது திறமையான எதிர்பார்ப்புகளுக்கு ஏற்ற மொழிபெயர்ப்பு மென்பொருள்கள் உருவாக்கப்படும். அதுவரை மொழிபெயர்ப்பிற்கு உதவக்கூடிய வரை நிலையிலான மொழிபெயர்ப்பையே கணிப்பொறி வழி கிடைக்கப் பெற இயலும்.

1.4. இயற்கை மொழி ஆய்வு

கணிப்பொறி ஒரு மொழியாய்வுக் கருவி. செய்திகளை 1,0 என்ற எண்களால் ஆய்வது தான் கணிப்பொறி. மனித மூளை 0,1 என்ற அடிப்படை எண்களை வைத்துக் கொண்டு எந்திர மொழி மூலம் சிக்கலான கணிப்பொறியை உருவாக்கியது. எழுத்துக்களும், எண்களும் தனித்தன்மையான குறியீடுகளும் (special symbals) 0,1 என்பவைகளால் தான் உருப்படுத்தம் செய்யப்படுகின்றன. உயர்நிலை மொழிகள் பொருள்கோளின் (interperters) ஒன்று கூட்டிகள் (assemblers) என்பன மொழியியலால் பயன்பெற்றன. இப்பொழுது மொழியியலார் கணினியை

அவர்களின் கடினமான செயல்பாடுகளைச் செய்யும் உண்மையான கருவியாகப் பயன்படுத்தும் காலம் வந்துவிட்டது. மொழியியல் சார் கணிப்பொறி ஆய்வை கணினி மொழியியல் ஆய்வு அல்லது இயற்கை மொழி ஆய்வு என்பர். இது பேச்சைத் தெரிந்துகொள்ளுதல், பேச்சை உருவாக்குதல், உரைகளை/உரைகளை பேசாக்குதல், பேச்சுக்களைப் உரை ஆக்குதல், இயந்திர மொழிபெயர்ப்பு, விரிதரவுகளை ஆய்தல், உரைகளை ஆய்தல், மொழி அலகுகளைக் பகுப்பாய்வு செய்தல் என்பனவற்றை உள்ளடக்கும். மொழியாய்வு மொழியைப் புரிந்துக் கொள்ளுதல் என்பதுடன் பின்வருவனவற்றையும் உள்ளடக்கும். உரை அல்லது பேச்சுச் செய்திகளை பொருள் கொண்டுப் புரிந்துக் கொள்ளுதல். மொழிபெயர்த்தல், உரைகள் அல்லது பேச்சுச் செய்திகளை உருவாக்குதல்.

1.4.1. இயற்கை மொழியும் செயற்கைமொழியும்

மனிதர்கள் இயற்கையாகப் பயன்படுத்தும் எந்த மொழியும் இயற்கை மொழியாகும். செயற்கையாக மனிதனால் உருவாக்கப்பட்ட மொழிகள் இயற்கை மொழியில் இருந்து வேறுபட்டவை. எடுத்துக்காட்டாக, கணிப்பொறியில் பயன்படுத்தும் (programming) மொழிகளான போர்ட்ரான் (Fortran) பேசிக் (Basic), சி (C), புரோலாக் (Prolog) லிஸ்ப், (Lisp) பெர்ல் (Perl) என்பவைகளெல்லாம் செயற்கை மொழிகளாகும்.

இயற்கை மொழியாய்வு (natural language processing) கணிப்பொறியைக் கொண்டு மொழியை ஆய்வு செய்யும் எல்லா நடவடிக்கைகளையும் உள்ளடக்கும். பேச்சையும் (speech) எழுத்து மொழியையும் ஆயும் கணினி நுட்பங்கள் இதில் அடங்கும். பேச்சு உருவாக்கம் (speech synthesis) இயற்கை ஒலியான பேச்சை உருவாக்குவது நுட்ப அடிப்படையில் சிக்கலானதாகும். என்ன பேசப்படுகிறது என்பதைப் புரிந்து கொள்ள சரியான இசையோட்டம் போன்றவை தேவை.

1.4.2. பேச்சைத் தெரிந்து கொள்ளல்

இது அடிப்படையில் தொடர்ச்சியான ஒலி அலைகளை வேறுபட்ட சொற்களாகப் பகுத்தல், இயற்கைமொழியைப் புரிந்துக் கொள்ளுதல் (natural language understanding), தனித்தனிச் சொற்களிலிருந்து பொருண்மைக்குச் செல்லுதல், இயற்கை மொழியை உருவாக்குதல்,

(natural language generation), அனுமானிக்க இயலாத உள்ளீடுகளுக்குப் பொருத்தமான பதில்களை உருவாக்குதல் என்பனவற்றை உள்ளடக்கும்.

1.4.3 இயந்திர மொழிபெயர்ப்பு

இது ஒரு இயற்கை மொழியை இன்னொன்றாக மொழிப்பெயர்ப்பு செய்தலைக் குறிப்பிடும். பொதுவாக இயற்கை மொழியாய்வு பேச்சை உட்படுத்தாமல் தான் பயன்படுத்தப்படுகின்றது. பேச்சு மற்றும் இயற்கை மொழியாய்வு என்பதைப் பேச்சையும் பிற இயற்கை மொழியாய்வுகளையும் உள்ளடக்கப் பயன்படுத்தலாம்.

1.4.4. கணிப்பொறி வழி மொழிப்பகுப்பாய்வு

சிக்கலான மொழி பழக்கத்தில் ஈடுப்பட தேவையானவைகளை ஆறு வேறுப்பட்ட பாகுபாடுகளாகப் பகுக்கலாம். ஒலியனியல் பகுப்பாய்வு, உருபனியல் பகுப்பாய்வு, தொடரியல் பகுப்பாய்வு, பொருண்மையியல் பகுப்பாய்வு, பயன்வழியியல் பகுப்பாய்வு, கருத்தாடல் பகுப்பாய்வு என்பர். மொழியை இம்மட்டங்களில் பகுப்பதை கணிப்பொறி வழி மொழிப்பகுப்பாய்வு என்பர். இவை கணினி மொழியியலுக்கு அடிப்படை ஆய்வுகளாகும்.

ஒலியனியல் பகுப்பாய்வு: ஒலியனியல் பகுப்பாய்வு ஒலி அடையாள குறிகளிடமிருந்து அல்லது கேட்பு அடையாளக் குறிகளிடமிருந்து சொற்களைப் பிரித்தெடுக்கும், பேச்சைத் தெரிந்து கொள்ளும்.

ஒலிகள் → சொற்கள்
←
/b/+/ɔ:/ +t → bɔ:t/ ‘boat’
←

க்+ஆ+க்+அ+ம் → காலம்
←

உருபனியல் பகுப்பாய்வு: இது உருபன்களிலிருந்து சொற்களைப் பெறல் சொற்களை உருபன்பகளாகப் பகுத்தல் என்பனவற்றை உள்ளடக்கும்.

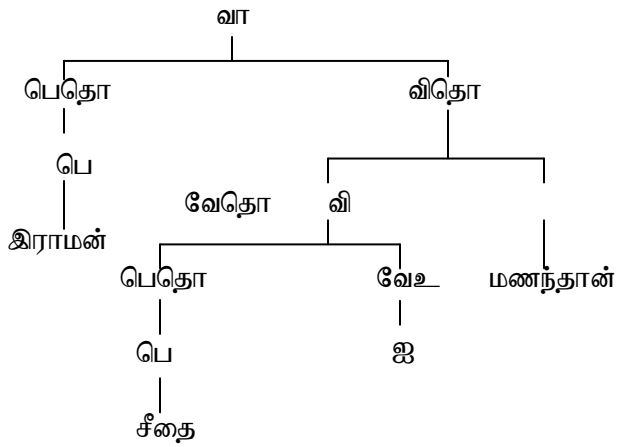
உருபன்கள் → சொற்கள்
←

மரம்+கள் → மரங்கள்
←

தொடரியல் பகுப்பாய்வு: இது சொற்களின் வரிசை அமைப்பிலிருந்து வாக்கிய அமைப்பைப்பெறல் அல்லது வாக்கியங்களைச் சொற்களாகப் பிரித்தல் என்பனவற்றை உள்ளடக்கும்.

சொல்வரிசை → வாக்கிய அமைப்பு
←

'இராமன் சீதையை மணந்தான்' என்ற சொல் வரிசையிலிருந்து பின் வரும் பகுத்தாயப் பட்ட கிளை அமைப்பைப் பெறலாம்.



பொருண்மையியல் பகுப்பாய்வு: வாக்கிய அமைப்பு சொற்பொருண்மை இவற்றிலிருந்து வாக்கியப் பொருண்மையைப் பெறல். எடுத்துக்காட்டாக, *இராமன் சீதையை மணந்தான்* என்று பகுத்தாயப்பட்ட கிளை அமைப்பிலிருந்து பின்வரும் தர்க்கக் கூற்றைப் பெறலாம்: மணந்தான் (இராமன், சீதை).

பயன்மொழியியல் பகுப்பாய்வு: வாக்கியப்பொருள் மற்றும் சூழல் இவற்றிலிருந்து சரியானப் பொருளைப்பெறல். இது பொது அறிவையும் முன்னர் வந்த கூற்றின் அறிவையும் உள்ளடக்கும்.

அவன் சீதையை மணந்தான் என்ற வாக்கியத்தில் வரும் பதிலீடுப் பெயர் (அவன்) முன்னர் வந்த வாக்கியத்தின் அறிவால் அவன் = ராமன் என்று புரிந்து கொள்ளப்படும்.

1.5. கணிப்பொறி வழி மொழி ஆய்வின் தேவைகளும் சிக்கல்களும்

கணிப்பொறியில் ஆங்கிலத்தைத் தவிர பிற மொழிகளையும் கையாளும் சூழ்நிலை வளர்ந்துள்ளது. குறிப்பாகக் கணிப்பொறியைத் தமிழில் கையாளுதல் கடந்த பத்து ஆய்டுகளுக்கு முன் தொடங்கப் பெற்று வளர்ச்சி அடைந்து வருகிறது. பொதுவான பயன்பாடுகளுக்குத் தமிழைக் கையாளுதல் கணிப்பொறியில் சமூக மேம்பாட்டை எய்த உதவுகிறது. இது தவிர மொழி ஆய்வுப் பணிகளுக்காகவும் கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

1.6. கணிப்பொறியில் மொழிகளைக் கையாளுதல் - வகைகள்

கணிப்பொறியில் மொழிகள் இரு விதமாகக் கையாளப் படுகின்றன:

அன்றாட பணிகளுக்காகக் கணிப்பொறியில் மொழியைக் கையாளுதல்
மொழி ஆய்வுக்காகக் கணிப்பொறியைக் கையாளுதல்

முதல் வகைப் பயன்பாட்டில் கணிப்பொறியில் மொழிகள் கையாளுவது கீழ்கண்ட பணிகளுக்கு உதவுகின்றன:

1. தமிழ்ச் சொல்லாய்விகளை உருவாக்குவதன் மூலம் கணிப்பொறியை ஒரு உயர்ரக மின்மத் தட்டச்சு போன்று தமிழுக்குப் பயன்படுத்துதல்
2. டி.டி.பி அதாவது மேஜைமேல் அச்சிடுதல் (Desk Top Printing (DTP)) போன்றவைகளுக்கான மென்பொருள்களைத் தமிழுக்கு உருவாக்கி அச்சுப் பணிகளுக்குக் கணிப்பொறியைப் பயன்படுத்தல்
3. நிறுவனங்களில் பணியாளர்கள் பட்டியல்கள், ஊதியப்பட்டியல்கள் மற்றும் வரவு செலவு அறிக்கைகளைத் தமிழில் தயாரித்தல்
4. வணிக நிறுவனங்களில் பொருள் வாங்கும் மற்றும் விற்கும் பட்டியல்களைக் கணிப்பொறி வழி தமிழில் தயாரித்தல்
5. தொலைவரி, பேஜர் போன்ற கருவிகளில் செய்தி தொடர்புக்குத் தமிழைக் கையாளுதல்

மேற்கண்ட பணிகளுக்காகக் கணிப்பொறியைக் கையயாளுவது தொடங்கப்பட்டுள்ளது. இத்தகைய வசதிகள் பொதுமக்களுக்குப் பயன்தரத்தக்க வகையில் அமைந்துள்ளன. எனவே இத்தகைய மென்பொருள்களை வடிவமைப்பதற்கு வணிக ரீதியில் செயல்படும் நிறுவனங்கள் முன் வருகின்றன. அரசும் ஆதரவு தருகிறது. இத்தகைய பணிகளுக்கான மென்பொருள் வடிவமைப்பதற்குத் தற்போது கணிப்பொறிகளில் உள்ள வசதிகள் போதுமானதாகும். மேலும் சாதாரண மொழி அறிவே வடிவமைப்பாளருக்குப் போதுமானதாகும்.

ஆனால் கணிப்பொறியை மொழி ஆய்விற்குப் பயன்படுத்தும்போது மொழியில் புலமையும் செறிவும் தேவைப்படுகின்றன. இத்தகைய பணிகளைக் கணிப்பொறி வல்லுநர்களும் மொழயியலாளர்களும் இலக்கியவாதிகளும் இணைந்து செயல்பட்டு வடிவமைக்கப்பட வேண்டும்.

கணிப்பொறியில் தமிழில் உள்ளீடு செய்வதற்கும் வசதிகள் உள்ளன. அதுபோல் வெளியீட்டைப் பெறுவதற்கு உயர்ரக அச்சுப்பொறிகள் உள்ளன. ஆனால் மொழி ஆய்வுப்பணிகளைச் செய்வதற்குப் போதுமான மென்பொருள்கள் இல்லை. இதற்கான முயற்சிகள் தொடங்கப்பட்டுள்ளன.

1.7. மொழி ஆய்விற்கு தேவைப்படும் கருவிகள்

மொழியாய்விற்குத் தேவைப்படும் கருவிகளாகப் பின்வருவனவற்றைப் பட்டியலிடலாம்: அகராதிகள், சொற்களஞ்சியங்கள், சொல்வங்கி, தகவல்தளம், கிளைப்பட வங்கி, பெருந்தரவு. எந்த ஒரு மொழியாய்வுப் பணியிலும் உயர்ந்த பட்சமாகச் சொல்லடைவு வரைக் கணிப்பொறியில் தயாரிக்கலாம். சொல்லடைவில் சொற்கள் சொல்வடிவம் என்ற நிலையில் உள்ளன. இதிலிருந்து அகராதிச் சொற்களைத் தயாரிப்பதற்கு உருபனியல் பகுப்பாய்வு தேவைப்படுகிறது. இப்பகுப்பாய்வினைச் செய்வதற்குச் செயல்முறைகள் கணிதக் கூறுகளாக வடிவமைக்கப்பட வேண்டும். இது எளிய பணியல்ல. இருப்பினும் இப்பணி மொழியாய்வின் தொடக்கப் பணியாக அமைகிறது.

தமிழில் அமைந்துள்ள செய்திகளைக் கணப்பொறியில் உள்ளீடு செய்வதற்கு விசைப்பலகை பயன்படுத்தப்படுகிறது. தமிழ் இலக்கிங்களையும் பிற படைப்புகளையும் கணிப்பொறியில் தட்டச்சு செய்து உள்ளீடு செய்ய நிறைய பொருட் செலவும் கால விரையமும்

ஆகும். எனவே அச்சு வடிவில் அமைந்த தமிழ்ச் செய்திகளையும் கையெழுத்து செய்திகளையும் கணிப்பொறி படித்து சொல் வடிவங்களாக மாற்ற ஒளி வருடி (Optical Character Recognizer (OCR)) என்ற தொழில் நுட்பம் பயனபடுத்தப்பட்டு ஒளி வருடிக்கான மென்பொருள்கள் உருவாக்கப்பட்டு வருகின்றன. இத்தொழில் நுட்பம் வெற்றி பெறும்பொழுது தமிழ்ச் செய்திகளை உள்ளீடு செய்வது மிக எளிய பணியாக அமையும்.

கணிப்பொறிக்குத் தரவேண்டிய கட்டளைகள் மனித மொழியில் தருவதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. ஒரு மனிதன் மற்றொருவரிடம் பேசுவது போன்றும் உரையாடுவது போன்றும் கணிப்பொறியுடன் உரையாடுவதற்கு மென்பொருள் உருவாக்கப்பட்டு வருகின்றன. இப்பணிகளெல்லாம் மொழியியலாளர்களின் மற்றும் கணிப்பொறியாளர்களின் கூட்டு முயற்சியால் நடைபெறுகின்றன.

தானியங்கி மொழி பெயர்ப்பிற்கு (automatic translation) ஒரு மொழியின் இலக்கண கூறுகள் அனைத்தையும் கணிப்பொறிக்குத் தர வேண்டி உள்ளது. அதாவது கணிப்பொறிக்கு மூலமொழி மற்றும் இலக்குமொழி இரண்டிலும் தேவையான அறிவினை ஊட்டவேண்டும். இப்பணியும் கணிப்பொறி தயாரிப்பாளரின் கூட்டுமுயற்சியிலே நடைபெறுகின்றன.

1.8. மொழியாய்வில் மொழியியலாளர்களுக்குத் தேவைப்படும் திறமைகள்

மொழி ஆய்வில் மொழியியலாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- கணிப்பொறியைப் பற்றிய அடிப்படை அறிவு
- மொழியின் இலக்கணம் மற்றும் அதனை கடினப் படுத்துவதில் திறன்
- ஒரு திட்டத்தில் எப்பணிகளைக் கணிப்பொறி வழி செய்ய இயலும் என்பதற்கான திட்டம்
- ஒன்றிணைந்து பணிபுரியும் கணிப்பொறியாளரிடம் மொழி சிக்கல் பற்றியும் தீர்பு பற்றியும் எடுத்துரைக்கும் திறன்
- சிக்கலுக்கான தீர்வினைப் படிப்படியாகப் பல்வேறு நிலைகளில் செயல்படுத்துவதற்கான திறன்
- கணிப்பொறி பயன்படுத்தும் போது தோன்றும் சிக்கல்களைச் சகமொழியாளர்களிடம் காட்டி அதற்குத் தீர்வு காணும் திறன்

1.9. மொழியாய்வில் கணிப்பொறியாளர்க்குத் தேவைப்படும் திறமைகள்

மொழியாய்வில் கணிப்பொறியாளர்களுக்குப் பின்வரும் திறமைகள் தேவைப்படுகின்றன:

- மொழியிலும் மொழி இலக்கணத்திலும் அடிப்படை அறிவு பெற்றிருத்தல்
- மொழியாய்வுப் பணியைப் பல்வேறு சிறு சிறு பணிகளாகப் பிரித்துக் கணிப்பொறி வழி செயல்படுத்துவதற்கான மென்பொருள்களை உருவாக்கும் திறன்
- மொழியாய்வுப் பணிகளுக்கு உதவும் கணிப்பொறி வழியமைப்புகளில் தேர்ச்சி பெற்றிருத்தல்
- மொழியாளருக்குத் தேவைப்படும் மாதிரி வழியமைப்புகளை உருவாக்கி மொழியாளருக்குச் செயல் விளக்கம் தந்து அவர்களின் ஒத்துழைப்பைப் பெறுதல்
- மொழியாளர் எண்ணும் பிரச்சனைகளுக்கான தீர்வுகளை உடனுக்குடன் எதிர்பாராமல் இருத்தல். மொழியாளர்களுக்குக் கணிதப்படுத்துவதில் பயிற்சியின்மையால் உடனுக்குடன் தீர்வினைத் தர இயலாது. இந்நிலையில் அவர்களுடன் ஒத்துழைத்து தீர்வு பெற முயற்சித்தல்
- மொழியாய்வுப் பணிகளுக்கான அவ்வப்பொழுது வெளிவரும் மென்பொருள் மற்றும் கணிப்பொறி வசதிகளைத் தெரிந்துக் கொள்ளுதல்

மேற்கூறியவாறு மொழியியல் ஆய்வுப் பணிகளுக்கு மொழியாளரும் கணிப்பொறியாளரும் சேர்ந்து ஒத்துழைக்கவேண்டும். கணிப்பொறி வழி மொழியாய்வுப் பணியினை மேற்கொள்வதற்கு தேவைப்படும் மென்பொருள்களை உருவாக்க கணிப்பொறியாளர்கள், மொழியாளர்கள், இலக்கியவாதிகள் ஆகியோரின் கூட்டு முயற்சி தேவையாகும்.

1.10. தமிழில் இயற்கை மொழி ஆய்வு

1.10.1. தமிழில் இயற்கை மொழி ஆய்வுகான முன்னேற்பாடுகள்

இந்திய மொழிகள் அனைத்தும் கணிப்பொறி அறிவியலில் உரிய இடத்தைப் பெறுவதற்காக நடுவண் அரசும் மாநில அரசும் பல்கலைக் கழகங்களும் ஆராய்ச்சி நிறுவனங்களும் தனியார் நிறுவனங்களும் பல்வேறு முயற்சிகளில் ஈடுபட்டுவருகின்றன. நடுவண் அரசின் தகவல் தொழில்நுட்ப அமைச்சரகமானது இதற்கென்றே ஒரு தனித் திட்டத்தை – இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technology Development of Indian Languages – TDIL) என்ற திட்டத்தை உருவாக்கிச் செயல்பட்டுவருகிறது. இதற்கென்று நாடெங்கும் 13 பல்கலைக்கழகங்கள் மற்றும் ஆய்வு நிறுவனங்களைத் தேர்ந்தெடுத்து, பல கோடி ரூபாய் நிதி உதவி அளித்து வருகிறது. இத்திட்டத்தின் கீழ் பல இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நடைபெற்று வருகின்றன.

கடந்த ஐந்தாண்டுகளில் தமிழகத்திலும் மாநில அரசானது தமிழ்மொழியின் தொழில்நுட்ப வளர்ச்சிக்காகப் பல்வேறு திட்டங்களைச் செயல்படுத்தியுள்ளது. தமிழ் மென்பொருள் வளர்ச்சிக்கான பல்வேறு திட்டங்களை அறிவித்துச் செயல்படுத்தி வருகிறது. இதனடிப்படையில்தான் தமிழ் இணையப் பல்கலைக்கழகமும் நிறுவப்பட்டது. தமிழ் இணைய மாநாடுகள் நடைபெற்று உள்ளன. தமிழ்மென்பொருள் உருவாக்கத்திற்காக மானிய உதவியைப் பலர் பெற்றுள்ளனர்.

சென்னைப் பல்கலைக்கழக மொழியியல் ஆய்வுப் பிரிவு, அண்ணாப்பல்கலைக்கழக கணிப்பொறியியல் துறை, சென்னை குரோம்பேட்டையில் அமைந்துள்ள AU–KBC மையம், அண்ணாமலைப் பல்கலைக்கழக மொழியியல் துறை, தஞ்சை தமிழ்ப்பல்கலைக்கழகம் மொழியியல் துறை, கோவை பாரதியார் பல்கலைக்கழக மொழியியல் துறை போன்ற பல்வேறு பல்கலைக்கழகத் துறைகளில் தமிழ் மொழியின் தொழில்நுட்ப வளர்ச்சிக்கான பல்வேறு ஆய்வுகள் நிகழ்த்தப்பட்டு வருகின்றன. தமிழில் மென்பொருள் தயாரிப்பாளர்கள் இணைந்து, கணித் தமிழ்ச்சங்கம் என்ற ஒரு அமைப்பைத் தொடங்கி பல்வேறு திட்டங்களை மேற்கொண்டு வருகின்றனர்.

தமிழ்மொழி தொழில்நுட்ப வளர்ச்சியின் நிலைப்பற்றி நாம் பார்ப்பதற்கு முன்னர், உலக அளவில் மொழி தொழில்நுட்பம் (Language Technology) என்ற வளர்ச்சிக் கட்டத்தில் உள்ளது என்பதைச் சற்றுப் பார்க்கலாம். மொழி தொழில்நுட்பம் என்று அழைக்கப்படுவது என்ன? உயிரித்தொழில்நுட்பம் (Bio-technology), வேதியியல் தொழில்நுட்பம் (Chemical Technology), தோல் தொழில்நுட்பம் (Leather Technology), தகவல் தொழில்நுட்பம் (Information Technology)

என்று பலத்துறைகளைப் பற்றி நம்மில் பலருக்கு அறிமுகம் இருக்கலாம். அந்த வரிசையில் தற்போது மொழித்தொழில்நுட்பம் (Language Technology) என்பதும் சேர்ந்துள்ளது. கணினித்துறையில் மனிதனின் இயற்கை மொழிகளின் பல்வேறு செயல்பாடுகளுக்கான மென்பொருள் மற்றும் வன்பொருள்களை உருவாக்கும் தொழில்நுட்பமே மொழித்தொழில்நுட்பம் என அழைக்கப்படுகிறது. எடுத்துக்காட்டாக, கணினியில் தமிழைத் தட்டச்சு செய்து, அதைச் சொல் திருத்தி, இலக்கணத் திருத்தி போன்றவற்றின் மூலம் செம்மைசெய்யும் சொல்லாளர் மென்பொருள் (Word Processor) மொழித்தொழில்நுட்பத்தின் ஒரு விளைபொருளாகும். அது போன்றே நாம் கணினியில் தட்டச்சுச் செய்வதைப் பேச்சாக மாற்றிக் கொடுக்கும் மென்பொருள் (Automatic speech Recognizer), தட்டச்சு செய்வதற்குப் பதிலாக நாம் கணினியில் பேசினால் அதை எழுத்து வடிவில் தட்டச்சு செய்து கொடுக்கும் மென்பொருள் (Text–To–Speech), இணைய தளத்தில் நமக்குத் தேவையானவற்றைத் தேடி எடுத்துக் கொடுக்கும் மென்பொருள் (Search Engine), இவையெல்லாம் மொழித் தொழில்நுட்பத்தின் விளைபொருட்களேயாகும். இவை போன்ற இயற்கை மொழிகளுக்கான மென்பொருள் மற்றும் வன்பொருட்களை உருவாக்கும் மொழித் தொழில்நுட்பத் துறையானது பல்துறை சார்ந்த ஒன்றாகும். கணினித் துறை, தகவல் தொடர்புத்துறை, உளவியல் துறை, மொழியியல் துறை ஆகிய நான்கு துறைகளைச் சேர்ந்த வல்லுநர்களும் இணைந்து செயல்படும் ஒரு துறை இந்தத் மொழித்தொழில்நுட்பத் துறையாகும்.

1.10.2. மொழித்தொழில் நுட்பம்

மேற்கூறிய ஒவ்வொரு துறையினருக்கும் மொழித்தொழில்நுட்பத்தில் ஈடுபடுவதற்கான நோக்கங்கள் வேறுபட்டிருக்கலாம். கணிப்பொறித்துறை வல்லுநர்கள் (computer scientists) கணிப்பொறியைச் செயற்கை அறிவுத் திறனுள்ள ஒரு கருவியாக (artificial intelligent machine) மாற்றும் நோக்கத்தில் மொழித் தொழில்நுட்பத்தில் ஈடுபடலாம். ஏனென்றால் மனிதனின் அறிவுத் திறனுக்கு அடிப்படையாக அவனது இயற்கை மொழி அமைவதால், கணிப்பொறிக்கு செயற்கை அறிவுத் திறனைக் கொடுக்க முயற்சிக்கும் கணிப்பொறித் துறை வல்லுநர்கள் முதலில் கணிப்பொறிக்கு இயற்கை மொழிகளைப் புரிந்து கொள்ளும் மற்றும் பயன்படுத்தும் திறனைக் கொடுக்க முயலுகின்றனர். அதற்காக அவர்கள் இயற்கை மொழிகளின் பண்பு மற்றும் அமைப்பு என்ன, நமது மனித மூளையானது எவ்வாறு இயற்கை மொழிகளைக் கற்றுக் கொள்கிறது, பயன்படுத்துகிறது போன்றவற்றைக் கண்டறிவதற்காக மொழித்தொழில்நுட்பம் மற்றும் மொழியியல் ஆய்வுகளில் ஈடுபடுகின்றனர்.

தகவல்தொடர்புத்துறை வல்லுநர்கள் (communication engineers) நாம் ஒரு முனையில் இயற்கை மொழிகள் வழியாகப் பேசும் போது அல்லது எழுதும்போது, அதை மின்னணுச்சாதனங்கள் எவ்வாறு எவ்விதச் சொற்சேதமும் தொடர்ச்சேதமும் கருத்துச்சேதமும் இன்றிக் கொண்டுசெல்கிறது அல்லது கொண்டுசெல்லாம் என்பதை ஆய்வதற்காக மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

உளவியல் நிபுணர்களோ (psychologists) நாம் மொழி வாயிலாக மற்றவர்களோடு கலந்துரையாடும்போது என்னென்ன உத்திகளைக் கையாளுகின்றோம். இந்த உத்திகளைக் கணிப்பொறி வாயிலாக இயற்கை மொழிகளைப் பயன்படுத்தும் போது எவ்வாறு பயன்படுத்தலாம் என்பதைக் கண்டறிவதற்காகத் மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

மொழியியல் வல்லுநர்கள் (Linguists) மொழிகளின் அமைப்பு என்ன, பண்பு என்ன, எவ்வாறு பேச்சொலிகளோ எழுத்துக்களோ கருத்து அல்லது பொருளை வெளிப்படுத்துகிறது என்பவற்றையெல்லாம் கண்டறிந்து, பின்னர் இதனடிப்படையில் கணிப்பொறி வாயிலாக இயற்கை மொழிப்பயன்பாட்டை மேற்கொள்வதற்காக மொழித்தொழில்நுட்ப ஆய்வில் ஈடுபடுகின்றனர்.

அடிப்படையில் மொழித் தொழில்நுட்பத்தின் ஆய்வுப்பொருள் நமது இயற்கை மொழிகளேயாகும். எனவே மொழி அல்லது மொழியியல் ஆய்வின் ஒரு பிரிவே மொழித்தொழில்நுட்பமாகும். எனவே தான் மொழியியலாளர்கள் கணினிமொழியியல் (Computational Linguistics) என்ற ஒரு மொழியியல் ஆய்வுப் பிரிவை உருவாக்கி வளர்த்து வருகின்றனர். அதன் ஒரு உட்பிரிவே மொழித் தொழில்நுட்பமாகும். மொழித்தொழில்நுட்ப ஆய்வில் அல்லது வளர்ச்சியில் மொழி ஆய்வாளர்களுக்கு மிக முக்கியப் பெரும்பங்குண்டு.

உலக அளவில் கணினிமொழியியல் மற்றும் அதன் உட்பிரிவான மொழித் தொழில்நுட்ப வளர்ச்சியின் பயனாகப் பல்வேறு மொழி மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஜெர்மனியில் ஆச்சிரியப்படத்தக்க ஒரு மொழித் தொழில் நுட்ப வளர்ச்சி அண்மைக் காலத்தில் ஏற்பட்டுள்ளது. அங்கு ஆங்கிலம் மட்டுமே தெரிந்த ஒருவர் சென்றால், அவர் தனது செல்போனில் ஆங்கிலத்திலேயே பேசலாம். மறுமுனையில் ஆங்கிலம் தெரியாத ஒரு ஜெர்மானியர் அதை ஜெர்மானிய மொழியில் கேட்கலாம். ஒரு ஜப்பானியர் தனது ஜப்பானிய

மொழியில் கேட்கலாம். அதுபோன்று ஒரு ஜெர்மானியர் அல்லது ஜப்பானியர் தங்களது மொழியில் பேசினால், மறுமுனையில் ஆங்கிலம் மட்டும் தெரிந்தவர் அதை ஆங்கிலத்தில் கேட்கலாம். அதாவது உடனடிக் கணினி மொழிபெயர்ப்பு என்ற ஒரு மொழித் தொழில்நுட்ப வளர்ச்சி அங்கு ஏற்பட்டுள்ளது.

கனடா நாட்டில் வானிலையைக் கண்டறியும் கருவிகள் தரும் வானிலை விவரங்கள் ஆங்கிலத்தில் தானாக வெளிப்படுத்தும் ஒரு மொழித் தொழில்நுட்பம் உருவாக்கப்பட்டு, தற்போது செயல்படுத்தப்பட்டுவருகிறது. அதுபோன்று விமானம் மற்றும் ரயில் பயணத்திற்கான பதிவுச் சீட்டுகளைப் பதிவு செய்யப்போகின்றவர்களிடம் மனித உதவியாளர்களுக்குப் பதிலாகக் கணிப்பொறியே உரையாடி அவர்களுக்குத் தேவையான பதிவுச் சீட்டைப் பதிவு செய்துதரும் மொழித் தொழில்நுட்ப வளர்ச்சி ஐரோப்பிய நாடுகளில் ஏற்பட்டுள்ளது.

1.10.3. தமிழ்மொழித் தொழில்நுட்ப வளர்ச்சியின் இன்றைய நிலை

நமது தமிழ்மொழித் தொழில்நுட்பம் எந்த வளர்ச்சிக் கட்டத்தை எட்டடியுள்ளது என்பதை இனிப் பார்க்கலாம். 15 ஆண்டுகளுக்கு முன் தமிழ்நாட்டில் கணிப்பொறியில் தமிழைப் பார்க்கவே முடியாத நிலை இருந்தது. கணிப்பொறியில் ஆங்கிலம் போன்ற மொழிகளில் மட்டுமே தட்டச்சு செய்யவோ அல்லது பிற கணிப்பொறிப் பயன்பாடுகளை மேற்கொள்ளவோ முடிந்தது. பின்னர் தமிழ் எழுத்துருக்களை (Tamil Fonts) கணிப்பொறியில் கொண்டு வரும் முயற்சி வெற்றிபெற்றது. அதன் பயனாகக் கணிப்பொறியில் தமிழைத் தட்டச்சு செய்யும் வளர்ச்சி ஏற்பட்டது. அதிலும் பல பிரச்சனைகள் நீடித்தன. தமிழுக்குப் பல வேறுப்பட்ட கணினித் தட்டச்சுப் பலகைகள் அறிமுகப்படுத்தப்பட்டன. மேலும் குறிப்பிட்ட ஒரு தமிழ் மென்பொருளில் தட்டச்சு செய்தலை மற்றொரு மென்பொருள் கொண்டு திருத்தவோ மாற்றவோ செய்ய இயலாத நிலை நீடித்தது. பின்னர் தமிழக அரசின் உதவியுடன் நடைப்பெற்ற தமிழ் இணையம் 99 மாநாடு ஓரளவுக்கு இப்பிரச்சனைகளைத் தீர்த்து வைத்தது. இதுபோன்ற முயற்சிகளினால் கணிப்பொறிக்கும் நமக்கும் இடையில் தமிழ்பயன்படத் தொடங்கியது. தமிழைக் கணிப்பொறியில் பார்க்கத் தொடங்கினோம். அதாவது பயன்பாட்டாளர் இடையூடகம் (User interface) என்ற அடிப்படையில் தமிழ் பயன்படத் தொடங்கியது. பல கணினி மென்பொருள்களில் தமிழ் எழுத்துக்கள் நடைபோடத் தொடங்கின. இந்த வளர்ச்சியைத் தமிழ் வாயிலாகக் கணிப்பொறியின் பயன்பாடு என்று அழைக்கலாம்.

ஆனால் ஆங்கில மொழிக்கு உருவாக்கப்பட்டுள்ள சொல்லாளர் (Word Processor), பேச்சறிவான் (Speech Recognizer), எழுத்து–பேச்சு மாற்றி (Text–To–Speech transferor), இணையத்தேடுவான் (Speech Engine) போன்ற பலமொழிசார் மென்பொருள்கள் தமிழுக்கு இன்று இல்லை. ஆங்கிலத்தைக் கணிப்பொறியில் தட்டச்சு செய்யும் போது, சொற்பிழை அல்லது இலக்கணப் பிழை ஏற்பட்டால் தானாக அவற்றைத் திருத்தித்தரும் சொல்லாளர்கள் பல உள்ளன. ஆனால் இன்னும் தமிழுக்கு ஒரு சரியான சொல்திருத்தியையும் இலக்கணத் திருத்தியையும் கொண்ட முழுமையான சொல்லாளர் உருவாக்கப்படவில்லை. இதுபோன்ற தமிழ் மென்பொருள்களைத் தமிழுக்கான மென்பொருட்கள் (Software For Tamil) என்று அழைக்கலாம். இவ்வகை மென்பொருட்கள் உருவாக்கப்படுவதில் என்ன தடைகள் என்பதைப் பின்னர் பார்க்கலாம்.

மருத்துவத்துறை மென்பொருட்கள் உருவாக்கப்படவேண்டுமென்றால், முதலில் குறிப்பிட்ட மருத்துவத்துறையில் அறிவு உடையோர் அத்துறையில் ஆராய்ச்சி மேற்கொண்டு, மருத்துவ உண்மைகளைத் தெளிவுப்படுத்தும்போதுதான், அவற்றை அடிப்படையாகக் கொண்டு, மருத்துவர்களின் உதவியோடு கணிப்பொறித்துறை வல்லுநர்கள் மருத்துவ மென்பொருட்களை உருவாக்கமுடியும். அவ்வாறு இல்லாமல் கணிப்பொறித்துறை வல்லுநர்களுக்கும் தான் உடம்பு, உயிர் இருக்கிறதே, அவர்களுக்கு அவைபற்றித் தெரியாதா, அவர்களே அதை அடிப்படையாகக் கொண்டு மருத்துவ மென்பொருட்களை உருவாக்கக்கூடாத என்று யாரும் கேட்பார்களா? கேட்க மாட்டார்கள். அதுபோல ஒரு மொழிக்கான மென்பொருளை உருவாக்கவேண்டும் என்றால், மொழியறிஞர்கள் அக்குறிப்பிட்ட மொழியின் அமைப்பு, பண்பு, செயல்பாடு போன்றவற்றை ஆராய்ந்து வெளிப்படுத்தும்போது தான், அவற்றை அடிப்படையாகக் கொண்டு கணிப்பொறித்துறை வல்லுநர்கள் அம்மொழிக்கு மென்பொருட்களை உருவாக்க இயலும். அவ்வாறு இல்லாமல் அம்மொழிபேசும் கணிப்பொறித்துறை வல்லுநர்கள் தாங்களே அம்மென்பொருட்களை உருவாக்க வேண்டியதுதானே என்று கேட்கக் கூடாது. அவ்வாறு கேட்பது தவறு. மொழி வல்லுனர்களின் துணையின்றி கணிப்பொறி வல்லுனர்கள் மொழிசார் மென்பொருள்களைச் செம்மையாக உருவாக்கவியலாது என்பது தான் உண்மை.

தமிழுக்கு ஒரு நல்ல சொல்திருத்தி (Spell Checker) உருவாக்கப்படவேண்டும் என்றால், தமிழ்ச் சொற்களின் பண்பு மற்றும் அமைப்பைப் பற்றி உண்மைகளைத் தெளிவாகத் தமிழ்மொழி

அறிஞர்கள் முன்வைக்கும்போதுதான், அதை அடிப்படையாகக் கொண்டு, கணிப்பொறித்துறை வல்லுநர்கள் சொல்திருத்தி மென்பொருளைத் தமிழுக்கு உருவாக்க முடியும். தமிழில் பெயர்ச்சொற்கள் எந்தெந்த விகுதிகளை எடுக்கும், எந்த வரிசையில் எடுக்கும், எவ்வாறு அவற்றை இணைத்து ஒரு முழுச் சொல்லை உருவாக்குவது, எவ்வாறு தமிழ்ச்சொற்களைப் பகுபத உறுப்புகளாகப் பிரிப்பது போன்ற உண்மைகளை எல்லாம் தெள்ளத்தெளிவாக, அறிவியல் பூர்வமாகத் தமிழ் மொழி அறிஞர்கள் எடுத்துச்சொன்னால், கணிப்பொறித்துறை வல்லுநர்கள் அவற்றை அடிப்படையாகக் கொண்டு சொல்திருத்தி மென்பொருட்களை உருவாக்க இயலும்.

இந்த இடத்தில் ஓர் உண்மையைத் தெளிவுப்படுத்த வேண்டியது அவசியம். மனித மூளை மொழியைப் புரிந்துகொள்ளும் முறை வேறு, கணிப்பொறி புரிந்துகொள்ளும்முறை வேறு. அதற்குக் காரணம், மனித மூளை ஒரு உயிர் வேதியியல் பொருள்; ஆனால் கணிப்பொறியில் உள்ள சிப் (chip) ஒரு மின்னணுப்பொருள். எனவே எந்த ஒரு அறிவையும் மனித மூளைக்குக் கொடுக்கும் முறை வேறு, மின்னணுச் சிப்புக்குக் கொடுக்கும் முறை வேறு, இவ்வேறுபாட்டை மனதில் கொண்டு மொழி அறிஞர்கள் தாங்கள் கண்டறிந்த உண்மைகளைக் கணிப்பொறிக்குப் புரியும் வகையில் மாற்றி அமைத்துகொடுத்தால் தான் கணிப்பொறி வல்லுநர்களின் வேலை எளிதாகும். இதுபோன்ற இலக்கணங்களைக் கணினி இலக்கணம் (Computational Grammar) என்று அழைப்பார்கள்.

ஆங்கிலம் போன்ற மொழிகளுக்குப் பல்வேறு நோக்கில் பல்வேறு கணினி இலக்கணங்கள் உருவாக்கப்பட்டுள்ளன. எனவேதான் அவற்றிற்கு நல்ல சொல்லாளர்கள், பேச்சறிவான்கள், எழுத்து–பேச்சு மாற்றிகள், இணையத் தேடுவான்கள், இயந்திர மொழிபெயர்ப்பிகள் போன்ற மென்பொருட்கள் நிலவுகின்றன.

இவ்வாறு கணிப்பொறி நோக்கில் மொழியாய்வு மேற்கொள்ளப்படுவதுதான் கணினி மொழியியல் என்று அழைக்கப்படுகிறது. அதன் உட்பிரிவே மொழித் தொழில்நுட்பம் ஆகும். எனவே தமிழ்க் கணினி மொழியியல் வளர்ச்சி அடையும்போதுதான் தமிழுக்கான பல மென்பொருட்கள் உருவாகும். இதுவே கணிப்பொறியில் உயர்நிலைத் தமிழ்மொழிப் பயன்பாடாகும். தற்போது தமிழ் எழுத்துருக்களை (Fonts) உருவாக்கி, கணிப்பொறியில் பிற மென்பொருட்களோடு உறவாடும் நிலைதான் (User Interface) ஏற்பட்டுள்ளது. அடுத்தக் கட்ட

நிலைக்கு நாம் செல்ல வேண்டும். இந்தி, தெலுங்கு, வங்காளம், மராத்தி போன்ற மொழிகளுக்கு இதுபோன்ற உயர்நிலை மென்பொருட்கள் உருவாக்கப்பட்டுள்ளன.

கணினிமொழியியலும் அதன் உட்பிரிவான மொழித் தொழில்நுட்பமும் இன்று உலகளவில் நன்கு வளர்ச்சி அடைந்துள்ளன. இதற்கென்று தனித்துறைகள் பல்வேறு பல்கலைக்கழகங்களில் உருவாக்கப்பட்டுச் செயல்பட்டுவருகின்றன. தமிழ்மொழியிலும் கணினிமொழியியல் கோட்பாடுகளின் அடிப்படையில் பல ஆய்வுகள் மேற்கொள்ளப்படவேண்டும். தமிழுக்குச் சொற்பகுப்பான் (Morphological Parser), சொல் உருவாக்கி (Word Generator), தொடர்பகுப்பான் (Syntactic Parser), தொடர் உருவாக்கி (Sentence Generator) போன்ற அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படுவதும் பௌதிக ஒலியியல் அடிப்படையிலான தமிழ்ப்பேச்சொலி ஆய்வு, பொருண்மையியல் (Semantics) அடிப்படையிலான ஆய்வு, விரிதரவு மொழியியல் ஆய்வு (Corpus Linguistics) போன்றவை மேற்கொள்ளப்படுவதும் கணினித் தமிழ் வளர்ச்சிக்கு இன்றியமையாதன. இவை இல்லாமல் தமிழுக்கு உயர்நிலையில் மென்பொருட்கள் உருவாக்க இயலாது.

கணிப்பொறிமயமாக உள்ள இன்றைய உலகில், உயர்நிலைத் தமிழ்மென்பொருட்கள் தமிழுக்கு உருவாக்கப்படாமல் உலக அளவில் மற்ற மொழிகளுக்கு இணையாகத் தமிழ் அடுத்த கட்ட வளர்ச்சியை எட்டவியலாது. அச்சு இயந்திரம் தோன்றியபோது, தமிழ் அதில் இடம் பெற்ற காரணத்தால்தான் இன்று தமிழ் எழுத்துலகில் நீடித்து வளர்ந்து வந்துள்ளது என்பதை நாம் மறந்துவிடலாகாது. அதுபோன்று கணிப்பொறியில் உரிய இடத்தைத் தமிழ் பெற்றால்தான் அடுத்த கட்ட வளர்ச்சிக்குத் தமிழ் செல்லும் என்பதை நாம் நினைவில்கொள்ள வேண்டும். தமிழுக்கு எழுத்துருவை உருவாக்கிக் கணிப்பொறியில் பார்ப்பதோடு நிறுத்திக்கொள்ளாமல், பல்வேறு உயர்நிலைத் தமிழ்மென்பொருள் உருவாக்கத்தில் நாம் கவனம் செலுத்தவேண்டும்.

எனவே, தமிழ் ஆசிரியர்களும் ஆய்வாளர்களும் மாணவர்களும் கணினிமொழியியல் மற்றும் மொழித் தொழில்நுட்பம் அடிப்படையில் தமிழ் ஆய்வை மேற்கொள்ளவேண்டும். உயர்நிலைத் தமிழ் மென்பொருட்கள் உருவாக்கத்திற்குத் தேவையான தமிழுக்கான அடிப்படை மொழித் தொழில்நுட்ப மென்பொருள் கருவிகள் உருவாக்கப்படவேண்டும்.

இயல்/பகுதி 2

கணிப்பொறி மூலம் தமிழ்மொழி ஆய்வு

2.1. மொழி ஆய்வும் கணிப்பொறியும்

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப்பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களை பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. வெறும் கணிக்கும் பொருளாக இருந்த கணிப்பொறி தரவுகள் சேமித்து வைக்கும் கருவியாக, தகவல் தொடர்பு கருவியாக தரவுகளை தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகிறது. மொழி செயல்பாட்டிற்கும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவில் உள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கானச் சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத்தட்டு, பதிவுநாடா போன்றவற்றில்) சேகரித்து வைத்துத் தேவையானபோது பயன்படுத்திக் கொள்ள இயலும்.

மொழிக்கான தரவுகள் பலவகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவோ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். இவைகள் உள்ளடங்கிய ஒரு பனுவலாகவும் (text) அமையலாம். வெவ்வேறு வகைப்பட்ட உரைகளிலிருந்து தெரிந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெரும்தரவு (corpus) என்று அழைக்கப்பெறும். தமிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக இன்றியமையாதப் பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய உரைகள், பல நடையியல் கூறுகளை உள்ளடக்கிய உரைகள், வட்டார வழக்கு, கிளை மொழி வழக்கு, சமுதாயமொழி வழக்கு, பேச்சு வழக்கு, இலக்கியவழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய உரைகள் எல்லாம் ஒருசேர பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம். 1. எழுத்து வடிவில் அமையும் பெருந்தரவு (Written corpus) மற்றொன்று பேச்சு வடிவில் அமையும் பெருந்தரவு (speech corpus). இத்தகைய பெருந்தரவுகள் மொழியாய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவிப் புரிகின்றன.

தரவுகள் கணிப்பொறியால் ஆயப்பட்டு தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன்தட்டு, குறுந்தட்டு, பதிவுநாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும்போது பயன்படுத்திக் கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையத்தளத்தில் பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல்மொழியைப் பயன்படுத்தி தேவையான தகவல்களைப் பெறுதலாகும். இதற்கு தேவை இயல்மொழி இடைமுகப்பு (Natural language Interface) இதனால் இயல்மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டு தேவையான தகவல்களைப் பெற இயலும். தமிழ்மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ்மொழியைப் பயன்படுத்தி தகவல்களைப் பெற உதவும். இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல்தேடும் பொறி (search engine) தகவல்களைப் பெற உதவும் முக்கியக் கருவியாகும். தமிழுக்கு இத்தகைய தகவல்தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு பெருமளவில் வெற்றியும் அடைந்துள்ளது.

கேள்வி-பதில் அமைப்புகள் கணிப்பொறியுடன் உரையாடும் அமைப்புகள் போன்றவை இயல்மொழி இடைமுகப்பை அடிப்படையாகக் கொண்டு அமைக்கப்பட்டிருந்தால் பாமரரும் பயன்படுத்தி தகவல்களைப் பெற இயலும். எடுத்துக்காட்டாக, தமிழ் இலக்கிய நூல் ஒன்றை விடுவிரலாக தந்து இயல்மொழி முகப்பு மூலம் தேவையான பதில்களை அல்லது தகவல்களை பெறும்படியாக கணிப்பொறியின் பயன்பாடு அமையலாம். விவசாயிகளுக்கான தரவு மையங்களிலிருந்து இயல்மொழியின் வழி வேண்டிய தகவல்களைப் பெறுதலும் இத்தகைய பயன்பாடாகும்.

மொழி உரைகளைப் படிநிலை (ஒலியனியலி நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழி அலகுகளாக பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துக் கொள்ள கணிப்பொறியைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக் கொள்வோம்:

> பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்

பள்ளி மாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும், *பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத்தொடர் என்றும், மேலும் *பள்ளி மாணவன், மரம், பலாப்பழம், ரகுமான்* என்பன பெயர்கள் என்றும், *பறி* என்பது வினை என்றும், *பெரிய* என்பது பெயரடை என்றும், *கவனமாக* என்பது வினையடை என்றும், *இலிருந்து* பின்னுருபு என்றும், *ஐ* என்பது வேற்றுமையுருபு என்றும், -*த்*- என்பது இறந்தகால இடைநிலை என்றும், *ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிபொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களை பகுத்தாய்வுச் செய்து தந்துவிடும்.

ஒரு காலகட்டத்தில் மொழியலார் கள ஆய்வு செய்து தகவலாளிகளின் உதவியால் ஆயப்படாத மொழிகளின் தரவுகளைச் சேகரித்து அதன் மீது சில மொழியியல் கொள்கைகளையும் கோட்பாடுகளையும் பயன்படுத்தி சில கருதுகோள்களின் அடிப்படையிலும் உற்றுநோக்கல் முலமும் அம்மொழிகளின் ஒலியனியல், உருபனியல், தொடரியல் அமைப்புகளைக் கண்டபிடிக்க முயலுவர். இது பண்புசார் மொழியியல் (qualitative linguistics) எனப்படும்.

இப்போது கணிப்பொறியின் வரவால் அத்தகைய நிலை மாறி மொழிகளின் விரிதரவுகள் உருவாக்கப்பட்டு அதன்மீது புள்ளியில் நெறிமுறைகளின் அடிப்படையில் மேற்சொன்ன மொழி அமைப்புகள் கண்டுபிடிக்கப்படும். இதை அளவுசார் மொழியியல் (quantitative linguistics) என்பர். தற்போது மொழி ஆய்வுகள் எல்லாம் கணினியைப் பயன்படுத்தி விரிதரவுகள் மீது செய்யப்படும் ஆய்வுகளாக மாறியுள்ளன.

மொழி உரையை இடு வரலாகத் (input) தந்து தக்க வழியமைப்புகளும் (programs) கட்டளைகளும் தந்தால் மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (out put) பெறலாம். கணிப்பொறியைக் கொண்டு மொழி உரைகளை வாக்கியங்களாகவும் வாக்கியங்களைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக்கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காக தொடரியல் ஆய்வி (Syntactic analyzer), உருபனியல் ஆய்வி (Morphological anayazer), ஒலியனியல் ஆய்வி (Phonological analyzer), பொருண்மையியல் ஆய்வி (Semantic anayazer) என்ற ஆய்விகளைத் தயாரிக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் தயாரிக்கப்பட்டுப் பயன்பாட்டில் உள்ளது. இவ்வாறு மொழிக் கூறுகளைப் கணிப்பொறி வழி பகுப்பாய்வது குறித்தும் தமிழில் இதற்கான முயற்சிகள் குறித்தும் இங்கு விளக்கப்படும். மேலும் இவ்வாய்விகள் அடிப்படையிலான மொழிசார் மென்பொருள்கள் அல்லது மொழிக் கருவிகள் உருவாக்குவது குறித்தும் இங்கு விளக்கப்படும்.

2.2. உரையிலிருந்து பேச்சு

பேச்சு சொல் பேச்சின் சிறு அலகுகளால் ஒருங்கிணைக்கப்பட்டது என்ற ஒலி அடிப்படையிலான எழுதும் ஒழுங்குமுறையில் உள்ளுறையும் கருத்து ஒலியனியலின் எல்லாத் தற்காலக் கோட்பாடுகளையும் உள்ளடக்கும் Ur-கோட்பாடு ஆகும். நாம் இங்கு தற்காலப் பேச்சறிதல் மற்றும் பேச்சுருவாக்கத் தொழில்நுட்பத்தையும் கணினி ஒலியனியல் (Computational Phonolgy) என்ற தொடர்புள்ள மொழியியலின் கிளைளையும் அறிந்துகொள்ள அடிப்படை உள்ளறிவையும் வழிமுறை வரைவுகளையும் அறிமுகப்படுத்த வேண்டும்.

தானியங்கு பேச்சறியும் முக்கியமான செயல்பாடு ஒலியியக்க அலைவடிவை உள்ளீடாக ஏற்றுச் சொற்களின் கோர்வையை வெளியீடாக உற்பத்தி செய்யும். எதிரிடையாகப் உரையிலிலிருந்து பேச்சுருவாக்கம் என்ற முக்கியச் செயல்பாடு உரை சொற்களின் கோர்வையை உள்ளீடாக ஏற்று ஒலியியக்க அலைகளை வெளியீடாகப் பெறுவதாகும். பேச்சைப் புரிந்து கொள்ளுதல் மற்றும் உருவாக்குதல் இவற்றின் பயன்பாடு பல்வகைப்பட்டதாகும்: தானியங்கு கேட்டெழுத்துதல் (dictation)/ ஒலிபெயர்ப்பு, கணிப்பொறிக்குப் பேச்சு அடிப்படையிலான இடைமுகங்கள் மற்றும் தொலை பேசிகள், பேச்சுக் குறையுள்ளவர்களுக்குக் குரல்சார்ந்த உள்ளீடு மற்றும் வெளியீடு என்பனவாகும்.

இங்கு நாம் பேச்சறிவான் மற்றம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள் என்பனவற்றில் கவனம் செலுத்துவோம்: எவ்வாறு சொற்கள் ஒலிகள் என்ற தனிப்பட்ட பேச்சு அலகின் அடிப்படையில் உச்சரிக்கப்படுகிறது? பேச்சறியும் ஒழுங்குமுறை அது புரிந்துக் கொள்ள இயலும் ஒவ்வொரு சொல்லுக்கும் உச்சரிப்பை வேண்டும். உரையிலிருந்து பேச்சு ஒழுங்குமுறை ஒவ்வொரு சொல்லுக்கும் ஒரு உச்சரிப்பை வேண்டும். நாம் இங்கு உச்சரிப்பை விளக்க ஒலி நெடுங்கணக்குகளை (phonetic alphabets) ஒலியில் களத்தின் பகுதியாக அறிமுகப்படுத்த வேண்டி வரும். மேலும் நாம் ஒலிப்புமுறை ஒலியியல் (articulatory phonetics) என்பதையும் அறிய வேண்டி வரும். இது எவ்வாறு வாயிலுள்ள ஒலிப்பான்களால் பேச்சொலிகள் உருவாக்கப்படுகிறது என்பதை ஆயும்.

ஒரு குறிப்பிட்ட ஒலி எப்பொழுதும் ஒவ்வொரு சூழலிலும் ஒரேமாதிரி உச்சரிக்கப்படுகிறது என்றிருக்குமானால் உச்சரிப்பை மாதிரிப்படுத்துவது எளிதாகும். ஆனால் இது அவ்வாறில்லை. எடுத்துக்காட்டாக, [த்] என்ற ஒலி வேறுப்பட்ட ஒலிச் சூழல்களில் மிக வேறுபட்டு மெய்ப்படுத்தம் செய்யப்படுகிறது. மொழியியலில் ஒலியனியல் (phonology) களம் வேறுபட்ட சூழலில் ஒலிகள் வேறுபட்டு மெய்படுத்தம் செய்யப்படுகின்றது என்பதையும் எவ்வாறு ஒலிகளின் இவ்வொழுங்குமுறை மீதி இலக்கணத்துடன் தொடர்புடையது என்பதையும் விளக்குகிறது. நாம் இவ்வேறுப்பட்ட மெய்ப்படுத்தங்களை விளக்க ஒலியனியல் விதிகளை எழுத வேண்டும்.

கணினி வழி ஒலியனியல் ஆய்வு கணினி ஒலியனியல் (computational phonology) என்று அறியப்படும். கணினி ஒலியனியலின் முக்கியமான பகுதி ஒலியனியல் விதிகளை

மாதிரிப்படுத்தும் கணினி இயங்குமுறைகளின் ஆய்வாகும். நாம் எவ்வாறு எழுத்துக்கூட்டல் மாற்றிகள் (spelling-rule transducers) ஒலியனியலை மாதிரிப்படுத்த இயலும் என்று காட்ட வேண்டும். பின்னர் ஒலியனியல் கற்றலின் (phonological learning) கணினி மாதிரிகளை விளக்க வேண்டும்: எவ்வாறு ஒலியனியல் விதிகள் இயந்திரக் கற்றல் வழிமுறைவரைவுகளால் தானாக ஊக்கப்படுத்த இயலும். இறுதியாக நாம் உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளில் ஒரு முக்கியமான சிக்கலுக்கு வேண்டி ஒலியனியலின் மாற்றி அடிப்படையிலான மாதிரியைப் (transducer-based model) பிரயோகிக்க வேண்டும்: எழுத்துக்களின் கோர்வைகளிலிருந்து ஒலிகளின் கோர்வைகளைப் பொருத்துதல். நாம் முதலில் பெரிய ஒலிப்பு அகராதியை உருவாக்குவதில் உள்ள சிக்கல்களை அளவிட வேண்டும். பின்னர் எவ்வாறு மாற்றி அடிப்படையிலான சொற்களஞ்சியங்களும் எழுத்துக்கூட்டல் அலகுகளும் எழுத்திலிருந்து உச்சரிப்புக்குப் பொருத்துவதற்கு உச்சரிப்புகளால் பெரிதுபடுத்தப்பட வேண்டும் என்று காட்ட வேண்டும்.

2.2.1. பேச்சு ஒலிகளும் ஒலி பெயர்ப்பும்

சொற்களின் உச்சரிப்பை ஆய்வது ஒலியியலின் ஒரு பகுதியாகும். ஒலியனியல் உலக மொழிகளில் பயன்படுத்தப்படும் பேச்சு ஒலிகளின் ஆய்வாகும். நாம் ஒரு சொல்லின் உச்சரிப்பை ஒலிகள் அல்லது கூறுகள் இவற்றை உருப்படுத்தம் செய்யும் குறியீடுகளின் கோர்வையாகும் என்று மாதிரிப்படுத்தலாம். நாம் ஒலிகளை ஆங்கிலம் போன்ற நெடுங்கணக்கு மொழியிலுள்ள எழுத்துக்களுடன் ஒற்றுமையுள்ள ஒலிக் குறியீடுகளால் உருப்படுத்தம் செய்வோம். எடுத்துக்காட்டாக, [ல்] என்பதால் உருப்படுத்தப்படும் ஒலி ல் என்ற எழுத்துடன் பொருந்தும். அதேபோல ப் என்பதால் உருப்படுத்தப்படும் ஒலி [ப்] என்ற ஒலியுடன் பொருந்தும். ஒலிகளுக்கு எழுத்துக்களைக் காட்டிலும் கூடுதல் வேறுபாடு இருக்கும். மீக்கூறு (prosody) இசைமை (pitch) மற்றும் மாத்திரை (duration) என்பனவற்றை உள்ளடக்கும்.

2.2.2. ஒலியன்களும் ஒலியன் விதிகளும்

ஒலிகள் வேறுபட்ட சூழல்களில் வேறுபட்டு உருவாக்கப்படும். ஒலியன்கள் தமக்குள் சூழல் அடிப்படையில் மாற்றொலிகளின் ஒரு குழுமத்தால் உருப்படுத்தம் செய்யப்படும். ஒலியனுக்கும் அதன் மாற்றொலியன்களுக்கும் இடையில் உள்ள உறவு ஒலியன்களின் விதியால்

பெறப்படும். எடுத்துக்காட்டாக, தமிழில் \க்\ என்ற ஒலியனை உருப்படுத்தம் செய்யப்படும் மூன்று மாற்றொலியன்களைப் பின்வருமாறு ஒலியனியல் விதியால் விளக்கலாம்:

k ⟶ g/N
k ⟶ x/V-V
k ⟶ k/ பிற சூழல்களில்

அதாவது மூக்கொலி முன் கடையண்ணக் குரலிலா ஒலியானது கடையண்ணக் குரல் ஒலியாகவும் இரு உயிர்களுக்கிடையில் கடையண்ண உரசொலியாகவும் பிற சுழல்களில் குரலிலா அடைப்பொலியாகவும் உருப்படுத்தம் பெறும்.

2.2.3. ஒலியனியல் விதிகளும் முற்றுநிலை மாற்றிகளும்

ஒலியனியல் விதிகளை முற்றுநிலை மாற்றிகளால் நடைமுறை படுத்தலாம். ஒலியனியல் விதிகளை மெய்ப்படுத்தம் செய்ய முற்றுநிலைத் தானியங்கியை பலவழிகளில் பயன்படுத்தும் கணினி ஒலியனியலின் வேறுப்பட்ட மாதிரிகள் இருக்கின்றன. இவற்றில் ஒன்று நாம் கணினி உருபனியலில் பார்க்க இருக்கும் கொஸ்கென்னியின் இருமட்ட உருபனியலாகும் (two-level morphology of Koskenniemi(1983)). இருமட்ட உருபனியல் சொல் வடிவுக்கும் புற வடிவுக்கும் இடையே உள்ள பொருத்தத்தின் மீது முற்றுநிலை நல்லுருவாக்கக் கட்டுப்பாடுகளாக மாதிரி செய்யும் உருபனியல்/ஒலியனியல் கோட்பாடாகும்.

கணினி ஒலியனியலின் உயர்வான சிக்கல்களை நேரிட ஒலியனியலின் இருமட்ட அல்லது முற்றுநிலை மாதிரியின் கூடுதல் சிறப்பான உபாயம் பயன்படுத்தப்படும். ஒலியனியல் விதிகளைப் பொறியால் கற்கச் செய்யலாம். ஒரு பரப்பிலிருந்து சிறிது தரவும் பிற தகவல்களும் தரப்படுகையில் அது அப்பரப்பிற்குத் தானாக ஒரு மாதிரியைத் தூண்டுவது பொறி கற்றல் ஒழுங்குமுறையின் வேலையாகும். இவ்வாறு ஒலியனியல் விதிகளைக் கற்கவேண்டி ஒரு ஒழுங்குமுறைக்கு அதிலிருந்து தூண்ட சொற்களின் (புற வடிவங்களின்) குறைந்தது ஒரு குழும விதிகள் தரப்படும். ஒரு கண்காணிக்கப்பட்ட வழிமுறைவரைவு (supervised algorithm) இத்தரவின் சிறுபகுதிக்கு விடைகள் தரப்படுகையில் இவ்விடைகளைப் பயன்படுத்தி அது முன்னர் பார்க்காத புதிய தரவைப் பொதுமையாக்கம் செய்யவியலும் ஒரு மாதிரியைத் தூண்டும். கண்காணிக்கப்படாத வழியமைப்பு (unsupervised algorithm) இதைத் தரவிலிருந்தே செய்கின்றது. கண்காணிக்கப்படாத வழியமைப்புகள் வகுப்புகளுக்குச் சரியான புலக்குறிகளைப் பார்க்காமல் இருந்தாலும், அவைகள் உருவாக்க வேண்டிய விதிகளின் அல்லது மாதிரிகளின்

இயல்பைப் பற்றி குறிப்பைத் தர இயலும். எடுத்துக்காட்டாக, தானியங்கி வடிவில் மாதிரிகள் இருக்கும் என்ற அறிவு ஒருவகைக் குறிப்பாகும். இம்மாதிரியான குறிப்புகள் கற்றல்சார் தவறானகருத்துகள் (learing bias) எனப்படும்.

2.2.4. உரையிலிருந்து பேச்சுக்குவேண்டி உரையை ஒலியன்களுக்குப் பொருத்துதல்

நாம் ஒரு மொழியின் ஒலிகளின் பட்டியலைத் தெரிந்துக் கொள்ளவும் எவ்வாறு ஒலியனியல் விதிகளை மாதிரிப் படுத்துவது என்பதை அறிந்துகொள்ளவும் செய்தால் எழுத்திலிருந்து அல்லது உரைச் சொல்லிருந்து அதன் உச்சரிப்புக்குப் பொருத்தும் சிக்கலை ஆயத் தயார் என்பதாகும்.

2.2.4.1. உச்சரிப்பு அகராதிகள்

உச்சரிப்பு அகராதிகள் உரையிலிருந்து பேச்சுக்குப் பொருத்த மிக பயனுள்ள கருவியாகும். எளிய உச்சரிப்பு அகராதிகள் சொற்களின் பட்டியலையும் அவற்றின் உச்சரிப்பையும் தரும். எ.கா.

Word	Pronunciation	Word	Pronunciation
cat	[kæt]	goose	[gus]
cats	[kæts]	geese	[gis]
pig	[pɪg]	hedgehog	[‘hɛʤ.hɔg]
pigs	[pɪgz]	hedgehogs	[‘hɛʤ.hɔgz]
Fox	[fax]		
Foxes	[‘fak.sɪz]		

PRONLEX, CMUdict, and CELEX என்ற இணையத்தள உச்சரிப்பு அகராதிகள் பொதுவாகப் பயன்படுத்தப்படுகின்றன. இவைப் பேச்சைத் தெரிந்து கொள்கைக்குப் பயன்படுத்தப்படுகிறது. இவற்றைப் பேச்சு கூட்டிணைப்புக்கும் பயன்படுத்த வேண்டி சுவீகரித்துக் கொள்ளலாம்.

2.2.4.2. அகராதியை நோக்குவதற்கு அப்பாற்பட்ட உரை ஆய்வு

ஆரம்பகால உரைகளிலிருந்து பேச்சாய்வுகள் உரையிலிருந்து ஒலிகளுக்குப் பொருத்த உச்சரிப்பு அகராதிகளை சார்ந்திருந்தன. முன்னர் கூறியபடி உரைகளிலிருந்து ஒலிகளாகப் பொருத்தமாற்றம் செய்யும் ஒரு வழி ஒவ்வொரு சொற்களின் உச்சரிப்பையும் உச்சரிப்பு அகராதியைப் பார்த்தறிந்து ஒலிகளின் கோர்வையை அகராதியிலிருந்து படிப்பதாகும். இந்த நெறிமுறை ஏற்கெனவே அகராதியில் தரப்பட்ட சொற்களுக்குப் பொருத்தும். ஆனால் பலச்சொற்கள் முன்னரே அகராதியில் தரப்படாததற்கான சூழல் சாத்தியம் இருக்கின்றன. இத்தகைய நிலையில் நாம் அகராதியில் தரப்படாத சொற்களின் உச்சரிப்பைத் தெரிந்துக்கொள்ள நிகழ்வுசாத்திய நுட்பங்களைப் (probabilistic techniques) பயன்படுத்த வேண்டிவரும்.

2.2.1.4.3. முற்றுநிலைமாற்றி அடிப்படையில் உச்சரிப்பு அகராதி

உரையிலிருந்து பேச்சு ஒழுங்குமுறைகளின் தொடக்கக்கால உச்சரிப்பு மாதிரி எழுத்திலிருந்து ஒவ்வொரு விதியும் எவ்வாறு ஒரு எழுத்து அல்லது எழுத்துக்களின் இணைப்பு ஒலிகளாக பொருத்தமாற்றம் செய்யப்பட்டன என்று குறிப்பிட்டுக் காட்டியது. கீழே விற்றன் (Witten, 1982) என்பவரின் விதி அடிப்படை வடிவு தரப்பட்டுள்ளது (Jurafsky and Martin, 2000: 125):

Fragment	Pronunciation
- p -	[p]
- ph -	[f]
- phe -	[fi]
- phes -	[fiz]
- place -	[pleɪs]
- placi -	[pleɪsi]
- plement -	[plɪmɛnt]

ஸ்.ப்ரோட் (Sproat) மற்றும் அவருடன் சேர்ந்த பிறரும் உரையிலிருந்து பேச்சுக்கு மாற்றிகளைப் பயன்படுத்தினார்கள். நாம் அவர்களின் மாற்றி அணுகுமுறையை ஐந்து பிரிவுகளாகப் பிரிக்கலாம்:

1. அகராதியிலுள்ள தனிச் சொற்களுக்கும் உருபன்களுக்கும் உச்சரிப்பை உருப்படுத்தம் செய்யும் முற்றுநிலை மாற்றி
2. உருபன்களின் சாத்தியமானத் தொடர்ச்சியை உருப்படுத்தம் செய்யும் முற்றுநிலைத் தானியங்கிகள்
3. ஒவ்வொரு உச்சரிப்பு விதிகளுக்கும் தனியான முற்றுநிலை மாற்றி
4. பெயர்களின் மற்றும் சுருக்கப்பெயர்களின் உச்சரிப்பை மாதிரிப்படுத்தப் பயன்படுத்தப்படும் உள்ளுணர்வு மற்றும் எழுத்திலிருந்து-ஒலி விதிகள்/ மாற்றிகள்
5. எந்த பிற அறியாத சொற்களுக்கும் வழுநிலை எழுத்திலிருந்து-ஒலி விதிகள்/மாற்றிகள்

உரையிலிருந்துப் பேச்சுக்கான உரையிலிருந்து ஒலிக்குப் பொருத்தமாற்றம் (Jurafsky and Martin, 2000:127):

Orthographic Lexicon	Lexicon
Regular Nouns	
cat	c/k a|æ t|t
fox	f|f o|ɑ x/ks
dog	d|d o|ɑ g|g
Irregular Singular Nouns	
goose	g|g oo|u s|s e|ɛ
Irregular Plural Nouns	
go:e o:e se	g| g oo|u: ee|i s|s e|ɛ

மேற்சொன்னப் படத்தில் அகராதிக்கு இரண்டு நிலைகள் இருக்கின்றன. ஒன்று உள்ளுறையும் சொல்சார் மட்டம்; மற்றொன்று இடைப்பட்ட மட்டம். இதில் இடைப்பட்ட மட்டத்தை புற மட்டத்திற்கு பொருத்தமாற்றம் செய்ய எழுத்துக்கூட்டல் விதிகளையும் உச்சரிப்பு விதிகளையும் பயன்படுத்தும் மாற்றிகளைச் சேர்க்க வேண்டும். இது வேறுபட்ட எழுத்துகூட்டல் விதிகளையும் ஒலியன் விதிகளையும் உள்ளடக்கும்.

சொற்களஞ்சியம் மற்றும் ஒலியனியல் விதிகள் மற்றும் வரிவடிவ விதிகள் என்பனவற்றைச் சொல் உருப்படுத்தத்திற்கும் புற உருப்படுத்தத்திற்கும் இடையில் பொருத்த மாற்றம் செய்யப்

பயன்படுத்தலாம். சொற்களஞ்சிய முற்று நிலை மாற்றிகள், அடிகள் மற்றும் ஒலியனியல் பண்புக் கூறுகளைக் கொண்ட சொல் மட்டத்திற்கும் உருபன்களின் எளிய இணைப்பை உருப்படுத்தம் செய்யும் இடைப்பட்ட மட்டத்திற்கும் இடையில் பொருத்தமாற்றம் செய்கிறது. பின்னர் ஒரு தனி எழுத்து விதிக் கட்டுப்பாடு அல்லது தனி ஒலியனியல் கட்டுப்பாடு இவற்றில் ஒன்றை உருப்படுத்தம் செய்யும் ஒருகுழும முற்றுநிலை மாற்றிகள் இடை மட்டம் மற்றும் புறமட்டம் இவற்றிற்கிடையில் பொருத்தமாற்றம் செய்ய இணையாகச் செயல்படுகின்றது. ஒவ்வொரு மட்டத்திற்கும் வரிவடிவ மற்றும் ஒலியனியல் உருப்படுத்தங்கள் இருக்கின்றன. உள்ளீடு சொல் வடிவில் இருக்கும் உரையிலிருந்து-பேச்சு பயன்பாடுகளுக்கு (எ.கா. ஒழுங்குமுறை சொல்லின் சொல்சார் அடையாளம், அதன் சொல்வகைப்பாடு, அதன் திரிபு இவற்றை அறியும் உரை ஆக்கத்திற்கு) முற்றுநிலை மாற்றிகள் சொல் வடிவிலிருந்து புற உச்சரிப்புக்குப் பொருத்த மாற்றம் செய்ய இயலும். உள்ளீடு புற எழுத்துக்கூட்டலாக இருக்கும் உரையிலிருந்து-பேச்சுப் பயன்பாடுகளுக்கு (எ.கா. "உரையை உரக்கப் படிக்கும்" பயன்பாடுகளுக்கு) முற்றுநிலை மாற்றிகள் உள்ளுறையும் சொல் வடிவு வழி புற வரிவடிவத்திலிருந்து புற உச்சரிப்பை பொருத்தமாற்றம் செய்யவியலும்.

இறுதி சிக்கல் தலைப்பெழுத்தன்களைப் (acronyms) பற்றியதாகும். தலைப்பெழுத்தன்களைக் கால அளவுடன் அல்லது கால அளவில்லாமல் எழுத்துக்கூட்டலாம். கால அளவுடன் உள்ள தலைப்பெழுத்தன்களை எழுத்துக்கூட்டி உச்சரிக்கலாம் (எ.கா [ai rɛs]). கால அளவின்றி தோன்றும் தலைப்பெழுத்தன்களைச் (AIDS, ANSI, ASCAP) சொல்லாக எழுத்துக்கூட்டிவோ உச்சரிக்கவோ செய்யலாம்; எனவே AIDS எனதை aid என்ற வினையின் படர்க்கைஇட வடிவு போன்றே உச்சரிக்கலாம். லிபர்மான் மற்றும் சர்ச் (Liberman and Church, 1992) சொற்களாக உச்சரிக்கப்படும் தலைப்பெழுத்தன்களின் ஒரு சிறிய அகராதியை வைத்திருக்கவும் மீதமுள்ளவற்றை எழுத்துக்கூட்டவும் யோசனை கூறுகின்றார்.

2.2.5. உரையிலிருந்து-பேச்சில் மீக்கூறு

முன்னர் கூறிய வரிவடிவிலிருந்து ஒலிக்குச் செல்லும் மற்றிச் செயல்பாங்கு பேச்சை உருவாக்கும் உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் பகுதிக்கு உள்ளீட்டின் முக்கியக் கூறை

உருவாக்கும். உள்ளீட்டின் மற்றொரு முக்கிய பாகம் மீக்கூறின் குறிப்பீடாகும். மீக்கூறு என்ற கலைச்சொல் பொதுவாகச் சொற்களஞ்சியத்திலிருந்து ஆக்கப்பட்ட ஒலிகளின் தொடர்ச்சியால் விளக்கப்படாத ஒரு வாக்கியத்தின் உச்சரிப்பின் நோக்குகளைக் குறிப்பிட பயன்படும். மீக்கூறு ஒலிகளைக் காட்டிலும் நீண்ட மொழி அலகுகளின் மீது செயல்படும்; எனவே இது மீக்கூறு நடப்பின் ஆய்வு என்று சிலநேரங்களில் அழைக்கப்படும்.

மீக்கூறுக்கு மூன்று முக்கிய ஒலியனியல் நோக்குகள் இருக்கின்றன: மேன்மை (prominence), அமைப்பு (structure) மற்றும் சுரம் (tune). மேன்மை அசையழுத்தத்தையும் (stress) ஒலியழுத்தத்தையும் (accent) உள்ளடக்கும். மேன்மை என்பது அசைகளின் தன்மையாகும்; இது ஒரு அசை மற்றொரு அசையைக்காட்டிலும் மேன்மையுடையது என்று ஒத்தறி அடிப்படையில் விளக்கப்படும். உச்சரிப்புச் சொற்களஞ்சியங்கள் சொல்சார் அசையழுத்தத்தைக் குறியீடுசெய்யும்; எடுத்துக்காட்டாக table என்பது அசையழுத்தத்தை அதன் முதல் அசையிலும் machine என்பது அதன் அசையழுத்தத்தை இரண்டாவது அசையிலும் கொண்டிருக்கும். there, the அல்லது a போன்ற செயல்பாட்டுச் சொற்கள் பொதுவாக ஒலியசையைக் கொண்டிருக்காது. சொற்கள் ஒன்றாக இணைக்கப்படுகையில், அவற்றின் ஒலியசையழுத்த அமைப்பொழுங்கு இணைந்து முழு வாக்கியத்திற்கும் ஒரு பொரிய ஒலியசையழுத்த அமைப்பொழுங்கை உருவாக்கும். எவ்வாறு ஒலியசையழுத்தங்கள் இணைகின்றன என்பதில் முரண்பாடு இருக்கின்றது. ஒலியசைகயின் இடம் கருத்தாடல் காரணிகளால் மிக வலுவாகப் பாதிக்கப்படும்.

சில சொற்கள் இயல்பாகவே ஒன்றாகக் குழுமுவதாகத் தெரிகின்றது என்ற அளவில் வாக்கியங்களுக்கு மீக்கூறு அமைப்பு இருக்கின்றது; சில சொற்களுக்கு அவற்றிற்கிடையில் காணத்தக்க இடைவெளி அல்லது பிரிநிலை இருக்கின்றது. பெரும்பாலும் மீக்கூறு அமைப்பு மீக்கூறு தொடராக்கம் (prosodic phrasing) அடிப்படையில் விளக்கப்படும். இப்பெரிய மீக்கூறு அலகுகளுக்குப் பயன்படுத்தப்படும் கலைச்சொற்கள் இசையோட்டத் தொடர் (ntonational phrase or IP) இசையோட்ட அலகு (intonational unit) மற்றும் சுர அலகு (tone unit) என்பனவற்றை உள்ளடக்கும்.

2.2.5.1. மீக்கூறின் ஒலியல் அல்லது ஒலியியக்கவியல் நோக்குகள்

மூன்று ஒலியனியல் காரணிகள் ஊடாட்டம் செய்யும் மற்றும் அவை பல வேறுபட்ட ஒலியியல் மற்றும் ஒலியக்கவியல் நடப்புகளால் மெய்ப்படுத்தம் செய்யப்படும். மேன்மையான அசைகள் மேன்மையற்ற அசைகளைவிட பொதுவாக ஓசையுடையதாகவும் நீண்டதாகவும் இருக்கும். மீக்கூறுத் தொடர் எல்லைகள் பெரும்பாலும் விட்டிசையாலும் எல்லைக்கு முன்னர் அசையின் நீட்சியாலும் சில சந்தர்ப்பங்களில் எல்லையில் இசைமையின் குறைப்பாலும் தொடாப்படும். இசையோட்டச் சுரம் அடிப்படை நிகழ்வெண் (F0) வெளிவரம்பில் வெளிப்படுத்தப்படுகின்றது.

2.2.5.2. பேச்சு உருவாக்கத்தில் மீக்கூறு

உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் ஒரு முக்கியமான வேலை மீக்கூறின் பொருத்தமான மொழி உருப்படுத்தத்தை உருவாக்குவாதாகும்; அவற்றிலிருந்து வெளியீட்டு பேச்சுப் அலைவடிவில் வெளிப்படுத்தப்படும் பொருத்தமான ஒலியிக்க அமைப்பொழுங்கை உருவாக்குவதாகும். இத்தகைய மீக்கூறுப் பகுதியைக்கொண்ட ஒரு உரையிலிருந்து-பேச்சு ஒழுங்குமுறையின் வெளியீடு ஒரு கால அளவும் F0 (இசைமை) மதிப்பும் உள்ள ஒலிகளின் தொடச்சியாகும். ஒவ்வொரு ஒலியின் கால அளவு ஒலியியல் சூழலைப் பொறுத்து அமையும்.

2.2.6. தமிழில் உரையிலிருந்து பேச்சு

தமிழில் உரையிலிருந்து பேச்சு உருவாக்கத்திற்கான முன்னேற்பாடுகள் பல நடந்துள்ளன. குறிப்பாக சென்னையிலுள்ள IIT-யில் இது தொடர்பான ஆய்வுகள் மேற்கொள்ளப்பட்டுள்ளன. அந்நிறுவனத்தில் பணிபுரிந்த யக்ஞநாராயணா, இராசேந்திரன் போன்றோர் உரையிலிருந்து பேச்சுத்தொடர்பான கருவிகளைச் செய்வதற்கு முயற்சிகள் மேற்கொண்டு வெற்றிபெற்றுள்ளனர். அண்ணாப்பல்கலைக்கழகத்தில் நடைபெற்ற RCTDIL-T என்ற ஆய்வுத்திட்டத்திலும் இதற்கான முயற்சிகள் செய்யப்பட்டுள்ளன. திருவனந்தபுரத்திலுள்ள உலகத் திராவிட மொழிகளின் பள்ளி (ISDL) மலையாளத்திற்கு உருவாக்கிய மலையாள உரையிலிருந்து பேச்சு ஒழுங்குமுறை தமிழ் உரைக்கும் நல்ல பேச்சு வெளியீட்டைத் தருகின்றது.

2.3. எழுத்துத்திருத்தியும் இலக்கணப்பிழைத் திருத்தியும்

சொல்லாய்வின் வழி ஆவணங்கள் (documents) உருவாக்கப் படுகின்றன. ஆவணம் ஒரு கடிதமாகவோ, சிறு குறிப்பாகவோ, கட்டுரையாகவோ, நூலாகவோ இருக்கலாம். சொல்லாய்வியில் ஆவணத்தை உருவாக்க பொதுவாக பின்வரும் முன்று முறைகள் பயன்படுத்தப்படுகின்றன:

1. கணிப்பொறியின் விசைப்பலகை (key board) வழி உள்ளீடு செய்தல்
2. பேச்சுணரி ஒழுங்குமுறையைப் (Voice Recognition System (VRS)) பயன்படுத்தி உரையை உள்ளீடு செய்தல்: இவ்வாறு செய்யும்பொழுது ஒலி வடிவில் பேசப்படும் செய்திகள் உரை வடிவிற்கு மாற்றப்படுகின்றன. இதற்குப் பேச்சிலிருந்து-உரை (Speech-to-Text (STT)) என்ற தொழில் நுட்பம் உதவுகிறது.
3. தட்டச்சு செய்தல் அல்லது அச்சடிக்கப்பட்ட உரையை ஒளி வழி எழுத்துணர்தல் (Optical Character Recongnition (OCR) என்ற தொழில் நுட்பத்தின் மூலம் கணிப்பொறிக்கு உள்ளீடு செய்து சொல்லாய்விற்கு ஆவணமாக மாற்றுவது.

உரையை உள்ளீடு செய்ய மேற்கண்ட முன்று முறைகளுள் ஏதேனும் ஒன்றைப் பின்பற்றலாம். பின்னர் சொல்லாய்வின் மூலம் அவைகளைக் கையாளலாம். உரைக்குள் எழுத்துப் பிழைகளும் இலக்கணப்பிழைகளும் இருப்பதற்கான வாய்ப்புகள் உள்ளன. ஆங்கில உரைகளில் உள்ள எழுத்துப்பிழைகளைக் கணிப்பொறி தானாகவே புரிந்துக்கொண்டு மாற்றுவதற்கு சொல்லாய்வில் வசதிகள் உள்ளன. இதற்கு எழுத்துப் பிழைகளைக் கண்டறிந்து திருத்துதல் என்று பெயர். இத்தகைய பிழைதிருத்தி தமிழ் சொல்லாய்வுகளில் உருவாக்கப்பட்டு வருகின்றன. எழுத்துப் பிழைகள் நிகழ்வதற்கான வாய்ப்புகள், எழுத்துப்பிழைகளின் வகைகள், பிழைகளைக் கண்டறிந்து திருத்துவதற்கான செயல்பாட்டு வழிமுறைகள் போன்றவற்றையெல்லாம் விரிவாக ஆராய்ந்து செயல்பட வேண்டி உள்ளது.

2.3.1. எழுத்துப் பிழை நிகழ்வதற்கான வாய்ப்புகள்

உரைச் செய்திகளைக் கீழ்கண்ட ஏதேனும் ஒருங்கு முறைகளில் உள்ளீடு செய்யப்படும்போது எழுத்துப்பிழைகள் நிகழ்வதற்கு வாய்ப்புகள் உள்ளன.

- விசைப்பலகை வழி உள்ளீடு செய்யும் பொழுது தட்டச்சுப் பிழைகள் (Typographical Errors) ஏற்படுகின்றன.
- பேச்சுணரி ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்ற பொழுது சொல்லின் உச்சரிப்பு மாற்றங்களினால் பிழைகள் ஏற்படுகின்றன. இத்தகைய பிழைகளுக்கு (Pronunciation Variation) என்று பெயர்.
- ஒளி எழுத்துணரி OCR வழி உள்ளீடு செய்கின்றபொழுது கருவிகளின் குறைபாடுகளினாலோ அல்லது எழுத்துக்களை சரியாக கருவிகள் படித்தறிதல் இல்லாததாலோ பிழைகள் தோன்றுகின்றன. உதாரணமாக ஆங்கிலத்தில் l என்ற எழுத்திற்கும் 1 என்ற எழுத்திற்கும் வேறுபாடு தெரியாததால் பிழை ஏற்படுகின்றன. பின்வரும் அட்டவணையில் சில சரியான எழுத்துக்களும் அவற்றிற்கிணையான தவறாகப் புரிந்துகொள்ளப்படும் எழுத்துக்களும் பட்டியலிடப்பட்டுள்ளன:

சரியன எழுத்து	தவறாகப் புரிந்துகொள்ளப்பட்ட எழுத்து
I	1
D	0
ri	n
0	∅

மேற்கண்டவாறு மூன்று வகையான உள்ளீட்டு முறைகளிலும் அதற்கான கருவிகளைப் பயன்படுத்தும் பொழுது பிழைகள் தோன்றுகின்றன. பேச்சறிதல் ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்ற பொழுது ஒரே வார்த்தையை அதன் இலக்கணக்கூறு மற்றும் தன்மைகளுக்கு ஏற்ப வெவ்வேறு விதமாக உச்சரிக்க வேண்டியுள்ளது. எடுத்துக்காட்டாக ஆங்கிலத்தில் object என்ற சொல் பெயர்ச்சொல்லாகவும் வினைச்சொல்லாகவும் பயன்படுகின்றது. பெயர்ச்சொல்லாகப் பயன்படும்போது ஆப்ஜிக்ட் ['ɔbʤikt] என்றும் வினைச்சொல்லாகப் பயன்படும்போது அப்ஜெக்ட் [əb'ʤekt] என்றும் உச்சரிப்பு செய்யவேண்டியுள்ளது. அதைப்போல் the என்ற சொல்லைச் சூழலுக்கு ஏற்ப தி [ði:] என்றும் த [ðə] என்றும் மாறுப்பட்ட இருவகைகளில் உச்சரிக்க வேண்டியுள்ளது. இவ்வாறாக, ஒரே சொல்லிற்கான உச்சரிப்பு மாற்றங்களுக்கு திட்டவட்டமாக நன்கு வரையறுக்கப் பட்ட விதிமுறைகள் உள்ளன. அவ்விதிமுறைகளைக் கணிப்பொறிக்கு ஏற்ற வண்ணம் வடிவமைப்புகளாக வடிவமைத்து இத்தகைய பிழைகள் நீக்கப்படுகின்றன.

2.3.2. பிழை நீக்கும் வழிமுறைகள்

உரைச் செய்திகளில் உள்ள எழுத்துப் பிழைகளை கண்டறிந்து மாற்றுவதற்கு இருவழிமுறைகள் பயன்படுத்தப்படுகின்றன:

1) பெய்சியன் விதிகள் (Bayesian Rules),
2) ஓசை ஊடக மாதிரி (Noisy channel model).

பெய்சியன் விதிமுறைகளைப் பின்பற்றும்பொழுது எழுத்துப் பிழைகளைக் கண்டறிவதற்கான வாய்ப்புகளை ஆராய்வதற்கு கணிதமுறையில் அமைந்த நிகழ்தகவுக் கோட்பாடு (Probability Theory) பயன்படுத்தப்படுகிறது. பெரும்பாலும் இம்முறைகள் தட்டச்சு செய்த உரைகளுக்குப் பயன்படுத்தப்படுகின்றன. இம்முறையை ஒளிவழி எழுத்துணரி (OCR) முறையிலும் பேச்சுப் புரிந்துகொள்ளல் ஒழுங்குமுறைகளிலும் (Voice recoginition system) உள்ளீடு செய்யும் உரைகளுக்கும் பயன்படுத்தலாம்.

இரண்டாவது ஓசை ஊடக மாதிரி (Noisy Channel Model) முறையில் உள்ளீடு செய்யப்படுகின்ற உரைகள் எவ்வாறு மாற்றமடைகின்றன, மாற்றமடைவதற்கான காரணங்கள் யாவை போன்றவைகள் ஆராயப்பட்டு ஒரு மாதிரி வடிவம் (Model) தயாரிக்கப்படுகின்றன. பின்னர் இம்மாதிரி வடிவத்திற்கான எதிர்மறைச் செயல்பாடுகள் (Reverse process) ஒரு மாதிரிப் படிவமாக உருவாக்கப்பட்டு பிழைகள் நீக்கப்படுகின்றன. எடுத்துக்காட்டாக, தொலைபேசியில் பேசுகின்றபொழுதும் ஒலிவாங்கி வழி பேசுகின்றபொழுதும் ஒருவர் பேசும் குரல் அமைப்புகள் மாறுபடுகின்றன. இத்தகைய கருவிகள் மூலம் கிடைக்கப் பெறுகின்ற முடிவுகள் ஒலி அமைப்பைப் பொறுத்துள்ளது; உள்ளபடியாகவே அமையாது. இந்நிலையில் அம்மாறுபாடுகள் நிகழ்வதற்கான காரணங்களை ஆராய்ந்து அதற்கு எதிர்மறையான திட்டங்களை வடிவமைக்கின்றபொழுது இவைகள் மூலம் நிகழும் பிழைகளை நீக்கலாம்.

2.3.3. எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் சிக்கல்களின் வகைகள்

எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் தொடர்பான சிக்கல்கள் பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

- மொழியில் இல்லாத சொற்களைக் கண்டறிதல் (Non-Word Error Detection)
- எழுத்துப்பிழை உள்ள செய்திகளை மட்டும் தனிமைப்படுத்தி ஆராய்தல்
- எழுத்துப்பிழை உள்ள சொல்லின் முன்பின் உள்ள அல்லது அச்சொல் இடம்பெற்ற தொடரின் பொருளை மையமாகக் கொண்டு பிழை நீக்கி சரிசெய்தல்

எழுத்துப்பிழைகளை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1. தட்டச்சுப்பிழைகள் (Typographical Errors)
2. சொல்லின் எழுத்து வடிவம் உரையாளருக்குத் தெரியாததால் ஏற்படும் பிழைகள்

தட்டச்சுப் பிழை பின்வரும் நான்கு வகைகளாகப் பிரிக்கப்படுகின்றன:

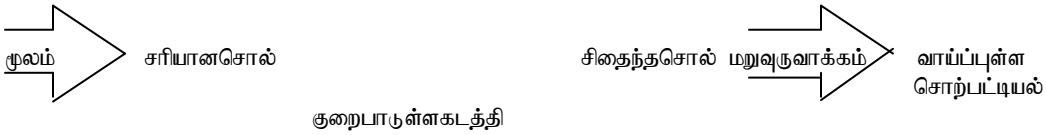
1. ஒரு எழுத்து விடுப்பட்டிருத்தல் (Delection Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் 'ட்' என்ற எழுத்தில்லாமல் 'சடம்' என இருப்பது விடுபடு பிழையாகும். இவ்வகையான பிழையை ஆங்கிலத்தில் 'Delection Error' என்பர்.
2. ஒரு எழுத்துக் கூடுதலாக இருத்தல் (Insertion Error): எடுத்துக்காட்டாக, *சட்டடம்* என்ற பிழையானச் சொல்லில் 'ட' என்ற எழுத்துக் கூடுதலாக இருப்பது செருகல் பிழையாகும். இவ்வகையான பிழைகளை ஆங்கிலத்தில் 'Insertion Error' எனபர்.
3. ஒரு எழுத்திற்குப் பதிலாக வேறொரு எழுத்தை மாற்றி அமைத்து விடுதல் (Substitution Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் முதல் எழுத்து 'ச' தவறாக அடிக்கப்பட்டும்போது அத்தவறான எழுத்து 'ச' ஆகவோ 'க' ஆகவோ 'ப' ஆகவோ இருக்கலாம். இம்மாதிரியான பதிலீட்டுப் பிழைகளை ஆங்கிலத்தில் (Substitution Error) என்பர்.
4. அடுத்தடுத்த இரண்டு எழுத்துக்கள் இடம் மாற்றி இருப்பது (Transposition Error): எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் பிழையாக 'ட' என்ற எழுத்தும் 'ட்' என்ற எழுத்தும் இடம் மாறி *சடட்ம்* என்று மாற்றி அடிக்கப்படுவது இடமாற்றத் தவறாகும். இவ்வகையானப் பிழைகளை ஆங்கிலத்தில் 'Transposition Error' என்பர்.

2.3.4. மொழியில் இல்லாதச் சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல்திட்டம்

ஒரு மொழியில் உள்ள சொல்லினைக் கண்டறிந்து சரிசெய்வதற்கு அகராதி பயன்படுத்தப்படுகிறது. கொடுக்கப்பட்ட சொல் அகராதியில் இல்லை எனில் அச்சொல்லை மொழியில் இல்லாத சொல் என கருதலாம். இதற்குத் தகுந்த அறிவுரையைக் கணிப்பொறி பயன்பாட்டாளருக்கு வழங்கும். இதற்காகப் பயன்படுத்தப்படும் கணிப்பொறிக்கான அகராதி சிறிதாக இருத்தல் நலம். மிகப்பெரிய அகராதியாக இருந்தால் அதிகமாகப் புழக்கத்தில் உள்ள எழுத்துப்பிழைச் சொற்கள் கூட அரிதாகப் பயன்படுத்தும் சரியான சொல்லாக இருக்கும்.

2.3.4.1. எழுத்துப்பிழைத் திருத்தி (Bayesian Medhod)

ஒரு உரையிலுள்ள எழுத்துப்பிழையைக் கண்டறிந்து திருத்துவதற்குப் பல செயல்திட்ட முறைகளை வடிவமைக்கப்படுகின்றன. அவற்றில் பெய்சியன் முறை நிகழ்தகவு அடிப்படையில் (Probability Thories) அமைந்ததாகும். இம்முறையின் அடிப்படைத் தத்துவம் பின்வரும் படத்தில் விளக்கப்பட்டுள்ளது:



உரைக்கான சொற்கள் தகுந்த உள்ளீட்டுக் கருவிகள் மூலம் உட்படுத்தப்படுகின்றன. உள்ளீடுக் கருவி கணிப்பொறியின் விசைப்பலகையாகவோ அல்லது ஒலியுணரி ஒழுங்குமுறையின் ஒலிவாங்கியாகவோ இருக்கலாம். தட்டச்சுச் செய்யும்பொழுது எழுத்துக்களின் சிதைவோ அல்லது ஒலியுணரி ஒழுங்குமுறையில் ஒலிச் செய்தியின் சிதைவோ ஏற்பட வாய்ப்புகள் உள்ளன. மூலக்கருவிகள் மூலம் உருவாக்கப்படும் சொல் உரிய கடத்திகள் (Channels) மூலம் கடத்தப்படுகிறது. கடத்திகளில் உள்ள குறைபாடுகள் காரணமாகவும் கடத்தும் செய்திகளில் சிதைவு ஏற்படலாம். இவ்வாறாகச் சிதைந்த சொல் மறு உருவாக்கம் (Decode) செய்யப்படுகிறது. இச்சொல் மூலக்கருவி மூலம் உருவாக்கப்பட்ட சொல்லுக்கு வாய்ப்புள்ள

சொல்லாகவே அமையும். இவ்வாறு சிதைந்த சொல்லுக்கு வாய்ப்புள்ள சொற்களின் பட்டியல் உருவாக்கப்படும். இப்பட்டியலிலிருந்து சிதைந்த சொல்கூறு திருத்தம் செய்யப்பெற்று பிழை நீக்கிய சரியானச் சொல் அடையாளம் காணப்படுகின்றது.

பேசியன் முறையில் எழுத்துத் திருத்திக்கான வழிமுறைகள் இருபெரும் கட்டங்களாக செய்யப்படுகின்றன:

1. பிழை நீங்கிச் சரியான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டடியல் தயாரித்தல்
2. இச்சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லுக்கும் மதிப்பெண் அளித்து அதிக மதிப்பெண் பெறும் சொல்லைப் பிழைத் திருத்தம் செய்யப்பட்ட சொல்லாகக் கருத வழிவகை செய்தல்

இம்முறையில் பிழை திருத்தம் என்பது சொல்லில் ஏதேனும் ஒரு எழுத்து மட்டும் பிழையாக இருக்கலாம் என்ற வாதத்தின் அடிப்படையில் செயல்முறைகள் மேற்கொள்ளப்படுகின்றன. எனவே பின்வரும் வகையிலான எழுத்து பிழைகள் மட்டும் இம்முறையில் எடுத்துக் கொள்ளப்படுகின்றன:

- சொல்லின் ஒரு எழுத்து விடுபட்டு இருத்தல்
- சொல்லில் ஒரு எழுத்து அதிகமாக இருத்தல்
- சொல்லில் ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து தவறாக இருத்தல்
- சொல்லில் அடுத்தடுத்த இரு எழுத்துக்கள் இடம் மாறி இருத்தல்

பேசியன் முறையில் எழுத்துத் திருத்தி உருவாக்குவதற்கு முதலில் பெருந்தரவு (Corpus) ஒன்றினைக் கணிப்பொறியில் உருவாக்கிக் கொள்ள வேண்டும். இந்த பெருந்தரவில் மொத்தம் ‘N’ சொற்கள் இருப்பதாகக் கெர்ளவோம். பிழையானச் சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் ஒன்றை உருவாக்க வேண்டும். இப்பட்டியலை உருவாக்குகின்றபொழுது ஒரு சொல்லில் ஒரு எழுத்தினால் மட்டும் பிழை நிகழ்வதாக எடுத்துக் கொள்ளப்படுகிறது. எனவே சொல்லில் (அ) ஒரு எழுத்து விடுபட்டு இருத்தல், (ஆ) ஒரு எழுத்து கூடுதலாக இருத்தல், (இ) ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருத்தல், (ஈ) அடுத்தடுத்த இரு எழுத்துக்கள் தவறாக இடம் மாறியிருத்தல் ஆகிய நான்கு மட்டும் கருத்தில் கொள்ளப்படுகின்றன. மேற்கண்ட

முறைகளின்படி பிழையான சொல்லிற்கு வாய்ப்புள்ள சொற்பட்டியலில் 'V' சொற்கள் இருப்பதாகக் கொள்வோம். இப்பட்டியலில் 'c' என்பதை ஏதாவது ஒரு சொல்லாகக் கொள்வோம். 'c(c)' என்பது 'c' என்ற சொல் பெருந்தரவில் எத்தனை முறை இடம்பெற்றிருக்கிறது என்பதைக் குறிப்பதாகக் கொள்வோம். இப்போது பிழை நீக்கிய சொல் 'c' என்ற சொல்லாக இருப்பதற்கான முன்நிகழ்வு எண்ணிக்கை பின்வரும் சூத்திரத்தின் மூலம் கணக்கிடப்படுகிறது:

$$P(c) = \frac{c(c)+0.5}{N+0.5V} - 1$$

$$P(\text{அடம்}) = \frac{c(\text{அடம்})+0.5}{N+0.5V}$$

(v = வாய்புள்ள சொல், C = விரிதரவு (corpus)

மேற்கண்ட சூத்திரம் ஒரு சொல் மட்டும் பெருந்தரவில் இடம்பெற்றுள்ள முன்நிகழ்வு (Prior) மதிப்பைத் தருகிறது. இரண்டாவது கட்டநிலையில் தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்களுக்கு மதிப்புகள் கண்டறியப்படுகின்றன. இதற்கு மாதிரி உரைகள் ஒரு தட்டச்சர் அல்லது பல தட்டச்சரைக் கொண்டு உள்ளீடு செய்யப்படுகின்றன. இவ்வுரையை அடிப்படையாகக் கொண்டு நான்கு விதமாக அணிகள் (Matrices) உருவாக்கப்படுகின்றன. தவறாகத் தட்டச்சு செய்த சொல் ஆங்கிலத்தில் 'typographical error' எனப்படுகிறது.

- del (x,y) என்பது xy என்ற இரு எழுத்துகளுக்குப் பதிலாக x மட்டும் எத்தனை முறை தவறாக அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
- ins (x,y) என்பது x என்ற எழுத்துக்களுக்குப் பதிலாக xy என்று தட்டச்சு செய்துள்ள எண்ணிக்கையைக் குறிக்கிறது.
- sub (x,y) என்பது x என்ற எழுத்திற்குப் பதிலாக எத்தனை முறை y என்ற எழுத்து தட்டச்சு செய்யப்பட்டுள்ளது என்பதைக் குறிக்கிறது.
- trans (x,y) என்ற எழுத்துக்களுக்குப் பதிலாக yx என இடமாறி அச்சடிக்கப் பட்டுள்ளது என்பதைக் குறிக்கிறது.

தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் இருந்து ஏதாவது ஒரு சொல் 'c' என்பதை எடுத்துக்கொள்வோம். 'c_p' என்பது 'c' என்ற சொல்லில் p-இடத்தில் உள்ள எழுத்தைக்

குறிப்பதாகக் கொள்வோம். இந்நிலையில் 'c' என்ற சொல் வாய்ப்பான சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

1. ஒரு எழுத்து விடப்பட்டிருக்கும் நிலையில்

$$P(t/c) = \frac{del[c_{p-1}, c_p]}{count\ [c_p - 1 c_p]}$$

2. ஒரு எழுத்து கூடுதலாக இருக்கும் நிலையில்

$$P(t/c) = \frac{ins\ [c_p - 1, t_p]}{count\ [c_{p-1}]}$$

3. ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருக்கும் நிலையில்

$$P(t/c) = \frac{trans\ [t_P, c_p]}{count\ [c_p]}$$

4. இரு எழுத்துக்கள் இடம் மாறி இருக்கும் நிலையில்

$$P(t/c) = \frac{trans[c_p, c_p+1]}{count[c_p, c_p+1]}$$

தவறான சொல் வாய்ப்புள்ள சொற்பட்டியலில் உள்ள 'c' என்ற சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

$$W = P(t/c) * P(c) - 3$$

இந்த சூத்தீரத்தில் P(c) என்பது c என்ற சொல் பெருந்தரவில் உள்ள முன் நிகழ்வு மதிப்பைத் தருகிறது. P(t/c) என்ற தவறாக தட்டச்சு செய்யப்பட்டு 'typo' என்ற சொல் 'c' என்ற சொல்லாக இருப்பதற்கு எவ்வளவு வாய்ப்புகள் உள்ளன என்ற மதிப்பைத் தருகின்றது. இவ்விரு எண்களின் பெருக்கல் தொகை, அதாவது P(t/c)□□□*P(c) என்பது இச்சொல்லிற்கான மதிப்பெண்ணைத் தருகிறது.

மேற்கண்டவாறு தவறான சொலலுக்கு வாய்ப்புள்ள சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லிற்கும் மேற்கண்ட மதிப்பினைக் கண்டறிய வேண்டும். எந்த செயால் அதிக மதிப்பெண் பெற்றிருக்கிறதோ அந்த சொல் அதிக சரியான சொல் என கருதப்படுகிறது. அதிக மதிப்பெண் பெற்றிருக்கும் சொல் பின்வரும் சூத்திரத்தின் மூலம் கூறப்பட்டுள்ளது.

$$W=\underset{c\epsilon V}{\operatorname{argmax}} \frac{P(t/c)P(c)}{P(t)}$$

மேற்கண்டவாறு அதிக மதிப்பெண் பெற்ற சொல் இடம் பெற்ற வாக்கியத்தின் பொருளுடன் ஒத்துப் பார்க்கப்படும். அது வார்த்தையின் பொருளுக்கேற்ற சொல் இல்லையெனில் அடுத்த மதிப்பெண் பெற்ற சொல் பொருத்திப்பார்க்கப்படும். இவ்வாறாகத் தவறாகத் தட்டச்சு செய்யப்பட்டுள்ள typo சரியான சொற்களாக மாற்றச் செயல் திட்டங்கள் உள்ளன.

இவ்வாறாக உரைகள் எழுதுதம்போது பிழைகளைக் கண்டறிந்து திருத்துவதற்குக் கணிப்பொறியில் பல செயல்திட்டங்கள் உருவாக்கப்படுகின்றன. அதில் மிகவும் பிரபலமானது மேற்கண்ட பேசியன் முறையாகும். இதுபோல் குறைந்த அளவு வித்தியாசம் உள்ள (minimum distance) என்ற மற்றொரு முறையும் பயன்படுகிறது.

2.4. உருபனியல் ஆய்வு

தமிழ்மொழி ஒரு திராவிட மொழி. இம்மொழி வினைச்சொல்லை இறுதியாகக் கொண்ட வாக்கியங்களை உடையதாகும். இம்மொழி இலக்கணச் செயற்பாங்குகளை (அதாவது இலக்கணப் பொருண்மைகளை) உருபனியலில் நிலையில் வெளப்படுத்தும் ஒட்டுக்கள் நிறைந்த மொழியாக விளங்குகிறது. இம்மொழியில் உள்ள ஒவ்வொரு வேர் உருபனும் ஆயிரக்கணக்கான சொற்களின் உருவாக்கத்திற்கு (அதாவது திரிபு மற்றும் ஆக்க வடிவுகளின் உருவாக்கத்திற்கு) அடிப்படையாக அமைகிறது. தமிழ் வாக்கியங்கள் எழுவாய்தொடரையும் வினையையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் செயப்படுபொருள் தொடரையும் வினையும் கொண்ட அமைப்பாகவோ எழுவாய்த்தொடரையும் பெயர்த்தொடரையும் (அதாவது பயனிலைத் தொடரையும்) கொண்ட அமைப்பாகவோ வரும். இதன் காரணமாகத் தமிழ் மொழி ஒரு SOV மொழி என்பர். தமிழ் மொழி ஒரு முன் மாதிரி SOV மொழிக்குள்ள எல்லாப் பண்புக்கூறுகளையும் பெற்றுள்ளது. பெயர்கள் வேற்றுமை உருபுகளுக்காகத் திரிபுறுகின்றன.

பெயர்த்தொடர்களுக்கும் வினைகளுக்கும் உள்ள உறவுகள் வேற்றுமை உருபுகளாலோ பின்னுருபுகளாலோ வெளிப்படுத்தப்படுகின்றன. வினைகள் கால இடைநிலைகளை ஏற்பதாகவும் எழுவாத்தொடருக்கு உடன்பாடாக வினைமுற்று விகுதிகளை ஏற்பதாகவும் வருகின்றன. இதன் காரணமாக எந்தக் கணிப்பொறி தொடர்பான மொழி ஆய்விற்கும் உருபனியல் பகுப்பாய்வு இன்றியமையாததாகவும் அடிப்படையாகவும் அமைகின்றது. மேலும் எழுத்துப்பிழை திருத்தி உருவாக்கத்திற்கும் இலக்கணப்பிழை திருத்தி உருவாக்கத்திற்கும் உருபனியல் பகுப்பாய்வு அடிப்படையாகும்.

உருபனியல் ஆய்வியன் உருவாக்கம் எந்த இயற்கை மொழி ஆய்வின் செயல்பாட்டிற்கும் மிக முக்கியமான முதற்படியாகும். வேறுபட்ட மொழி வகைகளுக்கு வேறுபட்ட ஆய்வு நெறிமுறைகள் இருக்கின்றன. எளிய இணைப்பு உருபனியல் உள்ள (simple concatinative morphology) மொழிகள் சொல்லும் செயற்பாங்கும் (item and process) மற்றும் சொல்லும் வரிசைமுறையும் (item and arrangement) என்ற மொழி மாதிரிகளால் உருவாக்கப்பட்ட ஆய்விகளால் தமது தேவையைப் பூர்த்தி செய்துகொண்டுள்ளன. பண்புக்கூறுகள் அடிப்படையிலான ஒருமைப்படுத்தும் கோட்பாடுகள் (unifivation theories) ஸ்பானிஷ் போன்ற மொழிகளுக்குப் பயன்படுத்தப்பட்டது. இக்காலத்தைய மிகப் புகழ்வாய்ந்த கணினிசார் இலக்கணமான தலைமையால் இயக்கப்படும் தொடரமைப்பு இலக்கணம் (Head-Driven Phrase Structure Grammar (HDPSG) பண்புக்கூறு அடிப்படையிலான உருபனியலைப் (Feature Basecd Morphology) பயன்படுத்துகின்றது. இங்கு தமிழ்மொழியின் உருபனியல் ஆய்வுக்குப் பயன்படுத்தப்பட்ட மாதிரிகளும் வேறு முன்னணி மாதிரிகளும் விளக்கப்படும்.

கணிப்பொறி வழி ஆய்வைப் பொறுத்தவரையில் இருமட்ட மாதிரி (two level model) உருபனியலின் முக்கிய கோட்பாடுகளில் ஒன்றாகும். இது பரந்த பரப்பெல்லையில்படும் மொழிகளுக்குச் செயல்படுத்தப்பட்டிருக்கின்றது. முற்றுநிலை தானியங்கி (Finite State Automata (FSA)) மாதிரி சென்னையிலுள்ள AUKBC ஆய்வு மையத்தால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வில் (API Morphological Analyzer) பயன்படுத்தப்பட்டுள்ளது. தலைமைத் தொடரில் அமைப்பு இலக்கணம் (Head Phrase Structure Grammar (HPSC)) தமிழுக்கு உமாராணியால் (Umarani, 2001) பயன்படுத்தப்பட்டுள்ளது. ஆம்பிள் என்ற மற்றொரு மாதிரி சொல்லும் வரிசையும் என்ற மாதிரியுடன் தொடர்புடையது.

பகுத்துலும் உருவாக்கலும் (Parsing and Generation) இயற்கை மொழிகளின் உருபனியல் ஆய்வின் இரு முக்கியமான நேர்வுகள் ஆகும். பகுப்பாய்வு என்பது பகுப்பாய்வி தரப்பட்ட உள்ளீட்டை புரிந்து கொள்ளவும் ஆயவும் செய்து பெரும்பாலும் அவற்றைப் பொருண்மை மற்றும் இலக்கணப் பண்புக் கூறுகளுடன் வெளியீடாகத் திருப்பித்தரும் செயல்பாடாகும். பல உரை மீட்பு ஒழுங்குமுறைகளுக்கும் (Text Retrieval Systems (TRS)) இயந்திர மொழிபெயர்ப்புச் (Machine Translation (MT)) செயல்பாடுகளுக்கும் பகுப்பாய்வு முக்கியமானது ஆகும். அதாவது இந்த இரண்டு செயல்பாடுகளுக்கும் தொடக்க இடம் பகுப்பாய்வாகும். ஒரு உணரி (recognizer) உள்ளீட்டுச் சொல்லை வேறுபட்ட உருபன்களாகப் பகுப்பாய்வு செய்கின்றது. ஆங்கிலச் சொல்லான reads என்பது பின்வருமாறு புரிந்து கொள்ளப்படும்: read என்ற வினையில் படர்க்கை இட ஒருமை எண் குறிப்பிடப்பட்ட நிகழ்கால வடிவம். பகுப்பாய்வி உணரியிலிருந்து அது புரிந்து கொள்ளும் சொற்களுக்குப் பொருத்தமான அமைப்பைத் தருகிறது என்ற அளவில் வேறுபடுகின்றது. இவ்வாறு ஒரு உருபனியல் பகுப்பாய்வி உள்ளீடு செய்யப்பட்ட சொல்லை அதன் உறுப்பு உருபன்களாகப் பிரித்து அதனைப் படிநிலை அமைப்பாகப் பகுப்பாய்வு செய்கிறது. இது சொல்லாக்கத்தில் உருபன்களை இணைப்பதற்கு மொழியில் குறிப்பிட்ட இயக்க நெறியின் ஒரு நோக்கை வெளிப்படுத்துகிறது. இவ்விக்கநெறி ஒரு வேருடன் கட்டுண்ட உருபுகளை முன்னொட்டாகவோ உள்ளொட்டாகவோ பின்னொட்டாகவோ வேருடன் சேர்க்கும். தமிழ்மொழியில் இது கூடுதலாகப் பின்னொட்டாக்கம் ஆகும். முன்னொட்டுக்கள் சமஸ்கிருதத்தில் இருந்து வந்த சில கடன் சொற்களில் இருக்கின்றன. மாறாக, உருபனியல் உருவாக்கி (Morphological Generator) உருபன்களை இணைத்துச் சொற்களை உருவாக்கும் நோக்கைப் பயன்படுத்துகிறது. இவ்வாறு உருபனியல் உருவாக்கி பகுப்பாய்வதன் மறுதலையாகும். இதன் உள்ளீடு ஒரு குழும உருபன்களோ அதன் இலக்கண அமைப்புகளோ ஆகும். இது அவற்றை ஒன்றிணைத்து ஒரு மொழியில் நல்வடிவாக்கமுள்ள சொற்களை உருவாக்குகிறது. உருபனியல் பகுப்பாய்வி இச்செயல்பாடுகளில் ஒன்றையோ அதற்கு மேலானவற்றையோ செய்கின்றது.

வேறுபட்ட உருபனியல் ஆய்விகள் அவை விளக்கும் மொழிகளின் அடிப்படையில் வேறுபட்ட கோட்பாடு அணுகுமுறைகளைப் பயன்படுத்துகிறது. அதாவது வேறுபட்ட மொழிகளின் இயல்பும் கலவைத்தன்மையும் அவற்றை ஆயப்பொருத்தமான கோட்பாடு அணுகுமுறைகளைத் தீர்மானிக்கிறது. வேறுபட்ட சட்டங்களின் அடிப்படையில் வேறுபட்ட

உருபனியல் ஆய்விகள் உருவாக்கப்படுகின்றன. வேறுபட்ட சட்டகங்களின் தேர்வும் அவற்றின் வேறுபட்ட நடைமுறைபடுத்தல்களும் ஒப்பீடு செய்யும் ஆய்வுக்கு அவசியமாகிறது.

2.4.1. உருபனியல் ஆய்விகளும் கோட்பாட்டு அணுகுமுறையும்

இப்பகுதியில் இருமட்ட உருபனியலைப் பயன்படுத்தும் பிசிகிம்மோ (PCKimmo), ஆம்பிள் (Ample) என்பனவும் சொல்லும் வரிசைமுறையும் என்ற மாதிரியின் அடிப்படையில் அமைந்த ஜிஎஸ்மார்ஃபும் (GSMorph) சொல்லும் செயற்பாங்கும் மாதிரியின் அடிப்படையில் அமைந்த இராமசாமியின் (Ramaswamy, 2000) தமிழ் உருவாக்கியும் (Generator for Tamil) சொல் ஒலியனியல் (Lexical Phonology) அணுகுமுறை அடிப்படையில் அமைந்த இரங்கநாதனின் (Renganathan,1997) டேக் தமிழும் (Tag Tamil) சொல்லும் வரிசைமுறையும் மற்றம் சொல்லும் செயற்பாங்கும் என்ற இரு மாதிரிகளையும் பயன்படுத்தும் வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும் இராசேந்திரனின் உருபனியல் ஆய்வியும் விளக்கப்படுகின்றன. இவை தவிர தமிழில் முயற்சிக்கப்பட்டுள்ள பிற உருபனியல் ஆய்விகளும் பட்டியலிடப்பட்டுள்ளன.

2.4.2. தமிழில் உருபனியல் ஆய்விகள்

தமிழில் கணிப்பொறி தொடர்பான செயல்பாடுகளுக்கு அடிப்படையாக அமைவது உருபனியல் பகுப்பாய்வும் உருவாக்கமும் ஆகும். பலர் இதில் பங்களிப்பு செய்துள்ளனர். பல உருபனியல் ஆய்விகள் தமிழில் உருவாக்கப்பட்டுள்ளன. இவற்றில் முக்கியமானவைகளைப்பற்றிய சிறு குறிப்பு கீழே தரப்பட்டுள்ளன.

2.4.2.1. இராஜேந்திரனின் தமிழ் உருபனியல் பகுப்பாய்வி

முதல் முதலில் இந்திய மொழிகளுக்கிடையில் இயந்திர மொழி பெயர்ப்பை மேற்கொள்ள வேண்டி அனுசாராகா (anusaraka 'language acessor') என்ற மொழிபெயர்புக் கருவி உருவாக்கத்திற்காகப் பாணினி வடிவவாதம் அடிப்படையில் ஆய்வு மேற்கொண்ட குழுவைச் சார்ந்த ஆய்வாளர்களால் DOE நிதி நல்கையில் தொடங்கிய ஆய்வுதான் இந்தியாவில் இயற்கை மொழி ஆய்விற்கு வித்திட்டது எனலாம். இந்திய மொழிகளுக்கிடையிலான இயந்திர

மொழிபெயர்ப்பு ஆய்வுக்காக உருவாக்கப்பட்ட உருபனியல் பகுப்பாய்விகள் அடிப்படையில் தமிழ்-இந்தி மொழி பெயர்ப்புக்காகத் தமிழ் உருபனியல் பகுப்பாய்வியும் இராசேந்திரன் உதவியுன் உருவாக்கப்பட்டது. AUKBC ஆய்வு நிறுவனக் குழுவினரால் உருவாக்கப்பட்ட அபி உருபனியல் பகுப்பாய்வியும் இராசேந்திரனின் மேற்பார்வையில் உருவாக்கப்பட்டதுதான். தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையில் பல்கலைக்கழக மானியக் குழுவின் நிதிநல்கையில் கீழ் மேற்கொள்ளப்பட்ட "Spell and Grammar Checher for Tamil" என்ற உயராய்வுக்காக PROLOG வழியமைப்பு மொழியைப் பயன்படுத்தி ஒரு உருபனியல் பகுப்பாய்வி இராசேந்திரனால் உருவாக்கப்பட்டது. இப்பகுப்பாய்வி இந்திய மொழிகளின் நடுவண் நிறுவன விரிதரவில் பயன்படுத்தப்பட்டு மதிப்பீடு செய்தபோது 95% வெற்றிகரமாக பகுத்துக்குறித்தல் வெளியீடு கிடைத்தது.

2.4.2.2. கணேசனின் உருபனியல் பகுப்பாய்வி

கணேசன் (Ganesan, 1994, 2003) அவர்கள் தமிழுக்காக இந்திய மொழிகளின் மையநிறுவனம் உருவாக்கிய மூன்று மில்லியன் சொற்கள் அடங்கிய விரிதரவை (CIIL Corpus) ஆய வேண்டி ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இவர் ஒலியனியல் மற்றும் உருபனியல் விதிகளைக் கையாண்டு இப்பகுப்பாய்வியை உருவாக்கினார். சமீபத்தில் சிறந்த தமிழ் உருபனியல் பகுப்பாய்வியை உருவாக்கியுள்ளார்.

2.4.2.3. கபிலனின் தமிழ் வினைகளுக்கான உருபனியல் பகுப்பாய்வி

கபிலன் (Kabilan, 1994) அவர்கள் தமிழ் வினைகளுக்கான ஒரு உருபனியல் பகுப்பாய்வியை உருவாக்கினார். இது பேராசிரியர் தெய்வசுந்தரத்தின் (மொழியியல் பகுதி, தமிழ்மொழித் துறை, சென்னைப் பல்கலைக்கழகம்) மேற்பார்வையின் கீழ் நடந்த தமிழ்சார் இயற்கை மொழி ஆய்வின் விளைவாகும்.

2.4.2.4. தெய்வசுந்தரத்தின் உருபனியல் பகுத்துக் குறிப்பான்

தெய்வசுந்தரம் தமிழுக்கென ஒரு சொல்லாய்வியைத் (Word Processor) தயாரித்துள்ளார். இது மிக நன்றாகச் செயல்பாடு செய்கின்றது. இது இருநிலை மாதிரியை (two-level-model)

அடிப்படையாகக் கொண்டது. புணர்ச்சி விதிகள் பகுப்பாய்வுக்குப் பயன்படுத்தப்படுகின்றன. இவர் ஆய்வி (Deiva sundaram and Gopal 2003) உருபொலியனியல் மற்றும் உருபனியல் என்ற இருநிலைகளில் செயல்படுகின்றது. இது ஒலியன் சேர்ப்பையும் (phonemic addition), ஒலியன் மாற்றத்தையும் (phonemic change) கணக்கிலெடுக்கின்றது. இவ்வாய்வி உருபனியல் அமைப்பொழுங்கு (mophotactics) அடிப்படையில் ஒட்டுகள் பகுதியுடன் இணைக்கப்படும் நிரலைக் கணக்கிலெடுத்துக் கொண்டு பகுப்பாய்வு செய்யும்.

2.4.2.5. AUKBC நிறுவனத் தமிழ் உருபனியல் பகுப்பாய்வி

AUKRC ஆய்வு நிறுவனம் (தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையுடன் இணைந்து) உருவாக்கிய API உருபனியல் சொல்லாய்வி தரவுக்கோப்பு, நிலை மாற்றமைவு அட்டவணை (State table.dat) என்பனவற்றில் தந்துள்ள தானியங்கி அட்டவணையைச் சார்ந்து இருக்கிறது. API உருபனியல் சொல்லாய்வி சொற்களை அதன் திரிபுகளிலிருந்து பிரித்தெடுக்கும்படி வடிவமைக்கப்பட்டுள்ளது.

2.4.2.6. வைஷ்ணவியின் தமிழ் உருபனியல் பகுப்பாய்வியும் உருவாக்கியும்

வைஷ்ணவி (Vaishnavi Ramaswamy, 2003) முனைவர் பட்ட ஆய்வுக்காகத் தமிழ் உருபனியல் பகுப்பாய்வை எடுத்துக்கொண்டு ஆய்வேட்டினைச் செய்துள்ளார். மேலும் வைஷ்ணவி (Vaishnavi Ramaswamy, 2002) தமது ஆய்வியல் நிறைஞர் ஆய்வேட்டிற்காக தமிழுக்கு ஒரு உருபனியல் உருவாக்கியை உருவாக்கி விளக்கமாகத் தந்துள்ளார். அவருடைய உருபனியல் உருவாக்கி சொல்லும் வரிசைமுறையும் மாதிரியையும் சொல்லும் செயற்பாங்கும் மாதிரியையும் பயன்படுத்தி செய்யப்பட்டுள்ளது. இந்த உருவாக்கி பிசிகிமோவின் (Pckimmo) கூட்டிணைப்பு நெறிமுறையால் செயல்படுகின்றது. வைஷ்ணவி இதைக் கலப்பு மாதிரி (hybrid model) என்று அழைக்கின்றார்.

PERL-இல் உருவாக்கப்பட்ட மற்றும் செயல்படுத்தப்பட்ட கலப்பு மாதிரி (hybrid model) உள்ளுருவாக்கப்பட்ட சொற்களஞ்சியத்தின் சட்டகத்தையும் மொழியின் வேறுபட்ட ஒவ்வொரு சொல் வகுப்பிற்கும் சொல்லும் வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் என்பனவற்றின் ஒரு குழும விதிகளையும் கொண்டிருக்கிறது. தரப்பட்ட உருபொலியனியல் விதிகளின்

அடிப்படையில் செயல்படும் இணையான துணை நடைமுறைகளை (corresponding subroutines) வேண்டுவதன் மூலம் இதன் ஆய்வு நிருபிக்கப்படுகிறது. இது ஒவ்வொரு வகைப்பாட்டின் சொல் வடிவுகளைக் கையாளும்படி வடிவமைக்கப் பட்டுள்ளது. இது முதன்மை அமைப்பின் கோப்புகளால் கட்டளையிட்ட படி செயல்படுகிறது. வழியமைப்பு மொழியில் மொழித் தகவலின் மரபுரிமை குறியாக்கம் இருக்கிறது. திறந்த இறுதியுள்ள உருபொலியனியல் மற்றும் மாற்றுருபு விதிகளின் (open ended morphophonimic and allomorphic rules) வழி மொழித் தகவலின் பகுப்பாய்வால் இது நடைப்பெறுகிறது (Vaishnavi Ramaswamy, 2003:101-110).

இம்மாதிரி உருபனியல் சொல்லும்வரிசைமுறையும் மற்றும் சொல்லும் செயற்பாங்கும் மாதிரிகளின் கலவையில் அதன் கோட்பாட்டு அடிப்படையை அமைத்துக் கொள்கிறது. செயற்பாட்டைச் சொற்பகுதியில் (word stem) உருபனியல் இயல்பைப் பொறுத்து ஒரு மாற்றத்தை (transformation) விவரணை செய்வதாக விளக்க இயலும். திரிபு நீக்கத்தின் (inflectional stripping) ஒவ்வவொரு நிலையிலும் குறிப்பிடப்பட்டுள்ள கட்டுப்பாடுகளின் முழுத்தொடர்சிகளின் வழி கடந்த பின்னர் ஒரு சொல் பகுப்பாய்வு செய்யப்படுகிறது.

ஒரு பகுப்பாய்வுத் தொகுதி (Analysis Module) வலமிருந்து இடமாகக் கோர்வை அமைவதை வருடிச் (scanning) செய்கிறது. ஒரு நேரத்தில் ஒரு பின்னொட்டை வருடி அறிந்து பிரித்தெடுக்கிறது. மற்றும் ஒவ்வொரு நேர்விலும் எடுத்துக்காட்டப்பட்ட ஒலியனியல் மற்றும் உருபொலியனியல் விதிகளின் உதவியுடன் மீதிச்சொல்லை மீட்டுருவாக்கம் செய்கிறது. கோர்வை தீர்வது வரை இது நடைபெறும். வெளியீடு கண்டுபிடிக்கப்பட்ட உருபன்களின் வடிவில் இருக்கும். பின்னர் அவை உறுதி செய்யப்படும்.

2.4.2.7. வின்ஸ்டன் குருசின் ஜி.எஸ். மார்ஃப் பகுப்பாய்வியும் உருவாக்கியும்

ஜி.எஸ். மார்ஃப் (GS Morph) என்று சுருக்கமாகக் குறிப்பிடபப்படும் எழுத்து அடிப்படையிலான அமைப்பு மொழியியலார் உருபனியல் பகுப்பாய்வி (Grapheme Oriented Structuralist Morphological Processor (GS Morph)) (Sengupta 1999) ஆம்பிள் போன்று சொல்லும் வரிசை முறையும் மாதிரியின் அடிப்படையில் அமைந்தது. இது ஒவ்வொரு உருபனின் விரிவான புற மெய்ப்படுத்தத்தை வேண்டும் அல்லது ஏற்படுத்தும். ஆம்பிள் ஒட்டுக்களை அவை பின்னொட்டுக்களா உள்ளொட்டுக்களா என்பதன் அடிப்படையில் வேறுப்பட்ட

கோப்புகளில் சேகரிக்க இயலும். ஆனால் ஜிஎஸ் மார்.்பின் பகுப்பாய்வு ஒரு வேர்ச் சொற்களஞ்சியத்தையும் (Root Lexicon) பின்னொட்டுக்களின் சொற்களஞ்சியத்தையும் (Suffix Lexicion) உட்படுத்தும்.

ஜிஎஸ் மார்.்ப் சொல்லடைவைப் பொருத்துவதால் செயல்படுகிறது. அதன் உள்ளீடு எழுத்து உருப்படுத்தத்தின் வடிவில் இருக்கிறது. அது வலது எல்லையிலிருந்து இடது எல்லைக்குப் பகுப்பாய்வைத் தொடங்குகிறது. முதலாவது அது தரப்பட்ட உள்ளீட்டைப் பூஜிய பின்னொட்டால் தொடரப்படும் வேராகக் கருதுகிறது. இதன்படி இது பின்னொட்டுச் சொற்களஞ்சியத்தில் பின்னொட்டுக்காகப் பரிசோதிக்கும். பின்னர் அது வேர்க்கோப்பில் பூஜிய ஒட்டிற்கு நேராக இருக்கும் சொல்லடைவுகளைப் (concordance) பார்க்கும். அது உள்ளீட்டைப் பொருத்துவதில் தோல்வியுற்றால் அது சொல் வடிவத்தைப் பற்றிய அதன் ஆரம்ப அனுமானத்தை மீளப்பெறும். அது உள்ளீட்டின் எழுத்துக்களை வெட்டி எடுக்கத் தொடங்கும். ஒவ்வொரு தடவையும் ஒரு எழுத்தை வெட்டி எடுக்கும்போது அது எல்லா வெட்டி எடுக்கப்பட்ட எழுத்துக்களும் சேர்ந்த ஒன்றைப் பின்னொட்டுச் சொற்களஞ்சித்தில் பொருத்தத்திற்காகப் பார்க்கும். வெட்டி எடுக்கும் செயற்பாங்கு முழுகோர்வையும் தீருவது வரை தொடர்ந்து நடைபெறும். பின்னொட்டுச் சொற்களஞ்சியத்தில் தேடல் வெற்றி பெற்றால் உள்ளீட்டின் மீதிப்பகுதி வேராக முன்மொழியப்பட்டு வேர்க்கோப்பில் பார்க்கப்படும். முன்மொழியப்பட்ட வேர் சொற்களஞ்சியத்தில் காணப்பட்டால் பின்னொட்டு மற்றும் வேர்க் கோப்புகளில் இருந்து அவற்றின் தொடர்புடைய பொருள் தரப்படும். இது வேர் மற்றும் பின்னொட்டுக் கோப்புகளைச் சொல்லடைவு செய்வதன் மூலமும் செய்யப்படும். சொல்லடைவு செய்யப்படும் சமயத்தில் வேர் மற்றும் பின்னொட்டுச் சொற்களஞ்சியங்களுக்குச் சிறு அட்டவணைகள் (hash tables) உருவாக்கப்படும். வழிமுறை வரைவுகள் (algorithms) தொடர்புடைய சிறு அட்டவணைகளில் தொடர்புடைய சொல்லடைவுகளைப் பொருத்திப் பார்க்க வேண்டி பரிசோதிக்கும். முழு உள்ளீடும் தீர்க்கப்படுவது வரை இச்செயற்பாங்கு நிற்பதில்லை என்பதை முக்கியமாகக் கவனிக்க வேண்டும். இப்படி அது ஒரு பொருத்தத்தைக் கண்டுபிடித்து வெளியீட்டைத் தந்தாலும், சாத்தியம் என்றால் பகுப்பாய்வி கூடுதலான பகுப்புகளைக் கண்டு பிடிக்க இயலும் என்பதாகும். இது வேர் மற்றும் மாற்றுருபுகளுக்கு ஒரே ஒலியனியல் வடிவு இருந்தால் பகுப்பாய்வு இரண்டிற்கும் எல்லாச் சாத்தியமான பகுப்பாய்வையும் தரும். தற்போதைய நடைமுறைப்படுத்தலில் விளைவு (effective) வேறுபாட்டைக் காலப் பின்னொட்டில் பார்க்கப்படும் நிலையில் தமிழில் வினைகளுக்கு ஒரு தனி வேர் உள்ளீட்டைத் தருவது, ஜி.எஸ். மார்.்பில் வேர் மற்றும் பின்னொட்டுக் கோப்புகள் அமைக்கப்படும் முறையின் காரணமாக இது

சாத்தியமாகிறது. இது பகுப்பாய்வின் போது ஒரே நடப்பை மிக எளிதாகப் பொருண்மை மயக்கம் தீர்க்க உதவுகிறது.

ஜி.எஸ் மார்.ப் பகுப்பாய்வி சொல்லின் வரிவடிவ உருபடுத்தத்தை உள்ளீடாக ஏற்கின்றது. சொற்கள் இடது எல்லையிலிருந்து பகுப்பாயப்படுகின்றன. முன்னர் விளக்கப் பட்டதுபோல் வேர் மற்றும் பின்னொட்டு சொற்களஞ்சியங்களின் பயன்பாட்டால் பகுப்பாய்வு வெற்றிகரமாக நடைபெறும்.

ஜி.எஸ். மார்.பின் உருவாக்கி உட்கூறு பகுப்பாய்வுக்கான அதே சொற்களஞ்சியங்களைப் பயன்படுத்துகின்றது. உள்ளீடு தரப்படும்போது உருவாக்கி அதை வேர்ச் சொற்களஞ்சியத்தில் பரிசோதித்து அதற்குத் தரப்பட்ட சொல்லடைவை எடுக்கும். அது பின்னர் அதே சொல்லடைவைக் கொண்ட கோர்வைகளைப் பின்னொட்டு அகராதியில் தேடும். ஏதாவது கோர்வை அதே சொல்லடைவுடன் இருந்தால் வேர்மாற்றுரு அப்பின்னொட்டுக் கோர்வையுடன் சேரும். இது பின்னொட்டில் உள்ள உருபன்களின் அர்த்தங்களுடன் திருப்பி அனுப்பப்படும். பகுப்பாய்வுக்குப் பயன்படுத்தப்படும் அதே கோப்புகள் பயன்படுத்தப்படுவதால் உருவாக்குதலும் துல்லியமாக இருக்கின்றது. (Winston Cruz, 2002:127).

வின்ஸ்டன் குருஸ் (Winston Cruz, 2000) தமிழ் வினைச்சொற்களைப் பகுப்பாய்வு செய்வதற்கு ஜி.எஸ். மார்.ப் முறையைக் கையாண்டார். ஜி.எஸ் மார்.ப் என்பது உருபன் அமைப்பொழுங்கு குறியீட்டால் ஆனது. வழிமுறை வரைவு (Algorithm) இரண்டு கோப்புகளை பொருத்தமானதா இல்லையா என்று மட்டுமே ஆய்வு செய்கிறது. சொல்லாய்வி இரண்டு வடிவங்களை மட்டுமே பயன்படுத்தி இரண்டு கோப்புகள் படைக்கப்பட்டுள்ளது.

துரைபாண்டியின் உருபனியல் பகுப்பாய்வு

தற்கால தமிழில் வினைச்சொல் வடிவத்தினை உருவாக்கும் பகுப்பாய்வு இயந்திரத்தை முழுமையாக உருவாக்கியுள்ளார்.

2.4.2.9. RCILTS-T-இன் தமிழ் உருபனியல் பகுப்பாய்வி

RCILTS-T (Resource Centre for Indian Language Technological Solutions-Tamil) என்ற அமைப்பின் தமிழ் உருபனியல் பகுப்பாய்வி அட்சரம் (atcharam) என அழைக்கப்படுகின்து. அட்சரம் வேர்ச் சொல்லிலிருந்து ஆக்கச் சொல்லைத் தனியாகப் பகுப்பாய்வு செய்து உருபனுடன் தொடர்புபடுத்துகிறது. இதனைப் பயன்படுத்தி 15 வகைகளை அடிப்படையாகக் கொண்டு இருபதாயிரம் வேர்ச்சொற்கள் தமிழ் அகராதியில் உருவாக்கப்பட்டுள்ளன. இது இரண்டு தொகுதிகளை (modules) உடையது. அதாவது பெயர் மற்றும் வினைச் சொல்லை அடிப்படையாகக் கொண்டது. இது ஊக விதிகளுடன் (heuristic rules) பொருண்மை மயக்கம் உடையதாகும். அதற்கு வினை மற்றும் பெயர் திரிபுகளைக் கையாண்டுள்ளனர்.

2.4.2.10. RCILTS-T –இன் உருபனியல் உருவாக்கி

இந்நிறுவனம் தமிழ்மொழியலும் உருபனியல் உருவாக்கியைத் தயார் செய்துள்ளது. அதற்கு அட்சயம் (Atchayam) எனப் பெயரிடப்பட்டுள்ளது. அட்சயம் என்பது பெயர் மற்றும் வினைச்சொற்களைக் கொண்டு இரண்டு முக்கிய உள்ளீடுகளைப் படைத்துள்ளது. இதில் பெயர்ச்சொல் பிரிவில் பின்னொட்டு இடம்பெறும். எடுத்துக்காட்டாக, பன்மை ஒட்டுக்களையும் திரிபு வடிவங்களையும் வேற்றுமை உருபுகளையும் பின்னுருபுகளையும் உருவாக்கியுள்ளனர். இதே போல் வினைப்பிரிவில் காலங்களையும் இடம்-எண்-பால் உருவாக்கிகளையும் பெயரச்சம் (relative), வினையெச்சம், பின்னொட்டு மற்றும் துணை வினைச்சொற்களையும் கொண்டுள்ளது. இது சந்தி விதிகள் மற்றும் 125 உருபனியல் விதிகளையும் பயன்படுத்தி உள்ளது. இது பெயரடை மற்றும் வினையடையையும் கையாண்டுள்ளது. இதில் சொல் மற்றும் வாக்கிய உருவாக்கியைக் கொண்டு ஆய்வு செய்யப்பட்டுள்ளது.

2.4.2.11. பொருண்மை மயக்கம் அல்லாத தமிழ் உருபனியல்

சொல் வடிவில் பொருண்மை மயக்கம் இருக்கும். இதற்கு மாறாக ஒப்புருச் சொல்லின் (homonymy) வகைப்பாடுகளில் எதிர் பாராத குறுக்கீடுகள் உருவாவதால் உருபனியல் பிறழ்ச்சி, ஒட்டுக்களின் பல்வகைச் செயல்பாடுகள் அல்லது நிச்சயமற்ற பின்னொட்டுச் சொற்கட்டுகளையும் (word bounding) உருவாக்குகிறது. இவ்விதம் வாக்கிய அமைப்புச் சூழல்களை இயல்பாக தீர்மானிக்கின்ற பகுப்பாய்வு ஆகும். இதனையே பொருண்மை மயக்கம் அல்லாத உருபனியல்

என்கிறோம். இரண்டு அடிப்படை அணுகுமுறைகள் சமவாய்ப்பு (probability) உடையது. விதியை அடிப்படையாகக் கொண்ட அணுகுமுறையில் சில பொருண்மை மயக்கம் தீர்வு காணப்பட முடியாமல் உள்ளது. ஆனால் மிகச் சில தவறுகளைக் கொண்டுள்ளது. பொதுவாகப் புள்ளியியல் அடையாளப்படுத்திகள் முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளைத் தருகின்றன. ஆனால் அவற்றறுள் நிறைய தவறுகள் காணப்படுகின்றன. சமவாய்ப்புள்ள முறையில் 1980-இருந்து உருபனியலில் பொருண்மை மயக்கமற்ற பகுப்பாய்வே அதிகம் இடம் பிடித்துள்ளது. நிலையான புள்ளியியல் முறைகளின் பயன்பாடு முழுவதும் பொருண்மை மயக்கமற்ற வெளியீடுகளாக உள்ளது. 'Influence of Morphology in Word Sense Disambiguation for Tamil' (தமிழ் உருபனியலில் பொருண்மை மயக்கம் அல்லாத சொல் செல்வாக்கு உடையது). பரிசோதனைகளை மேற்கொள்ளும்பொழுது மேற்பார்வை மற்றும் பகுதி மேற்பார்வை அணுகுமுறையை மேற்கொண்டு உருபனியல் திரிபுகளின் பாதிப்புமுறைகளை பற்றியும் திடமான தீர்வான உருபனியலில் பொருண்மை மயக்கம்மற்ற பகுப்பாய்வை தமிழ்மொழியிலும் பிற திரிபு மொழிகளிலும் பயன்படுத்தலாம் எனத் தீர்வு கண்டுள்ளனர்.

2.5. கணினி தொடரியல்

கணினி தொடரியல் ஆய்வு வாக்கியங்களைத் தொடர்களாகவும் சொற்களாகவும் பகுத்து அவற்றின் தொடரியல் மற்றும் சொல்வகைப்பாட்டுப் பண்புக்கூறுகளை அடையாளப்படுத்தித் தரும் செயல்பாடாகும். வாக்கியங்கள் தரப்படுகையில் அவற்றை தொடரியல் கூறுகளாகப் பகுப்பது மற்றும் தொடரியல் கூறுகள் தரப்படும் போது வாக்கியங்களை உருவாக்குவது என்ற இரு செயல்பாடுகள் இதில் அடங்கும்.

2.5.1. கணிப்பெறி வழி மொழி ஆய்வுக்கான இலக்கண வடிவமைப்புகள்

ஒரு கூற்றின் பொருத்தமான பொருளைத் தருமாறு சொற்களுக்கு இடையே உள்ள உறவுகளை பிரித்தெடுக்கும் நுட்பத்தைத் தருவது இந்த இலக்கண வடிவமைப்பின் முக்கியமான செயல்பாடாகும். வாக்கியங்களை பகுப்பாய்வு செய்வதற்குப் பயன்படுத்துகின்ற செயல்முறையிலிருந்து தனியாக ஒரு குழும விதிகளால் நாம் மொழியை பண்பாக்கம் செய்ய விரும்புகிறோம். அம்மாதிரியான குழும விதிகள் முறையான இலக்கணம் எனப்படும். முறையான இலக்கணம் ஒரு குழும இலக்கண வாக்கியங்களை வரையறுக்கிறது. அவ்வாக்கியங்களுக்கு

அமைப்பைத் தருகின்றது. மொழியைப் பேசுகின்றவர்களின் இலக்கணம் குறித்த உள்ளுணர்வுடன் பொருத்துகின்ற வாக்கியத்தின் பொருளை வரையறுக்கின்ற முறையான இலக்கணத்தை உருவாக்குகிறதுதான் நமது செயல். தொடக்ககால கட்டத்தில் சூழல்வரையறை இல்லாத இலக்கண வடிவவாதம் பயன்படுத்தப்பட்டது. தற்போது கட்டுப்பாடு அடிப்படையிலான இலக்கண முன்மாதிரிகள் பயன்படுத்தப்படுகின்றன. அவை கீழே பட்டியலிடப்பட்டுள்ளன:

- செயல்பாட்டு ஒருங்கிணைக்கப்பட்ட இலக்கணம் (Functional Unification Grammar)
- தலை இயக்கத் தொடரமைப்பு இலக்கணம் (Head-driven phrase Structure Grmmar)
- செயல்பாட்டு இலக்கணம் (Fuctional Grammar)
- சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar)
- வகைப்பாட்டு ஒன்றிணைக்கப்பட்ட இலக்கணம் (Categorical Unification Grammar)
- கிளை சேர்க்கப்படும் இலக்கணம் (Tree adjoining Grammar)

2.5.2. ஆழமில்லாப் பகுப்பாய்வு

நாம் ஆழமில்லாப் பகுப்பாய்வு என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளியீட்டைக் காட்டிலும் நிறைவுடைய ஆய்வைக் குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகின்றோம். ஒரு ஆழமில்லாப் பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அக அமைப்பையும் செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். ஆழமில்லாப் பகுப்பாய்வு மூன்று முன்மாதிரித் தொகுப்புகளை உள்ளடக்கியது:

1. சொல்வகை அடையாளப்படுத்தல் (parts-of-speech tagging): ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை) உறுதி செய்யும்
2. தொடர்பகுதியாகப் பிரித்தல் (chunking): இதில் சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்பும் தரப்படுகையில் எந்தச் சொற்கள் தொடர் பகுதிகளாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பித்தொடர் போன்றவைகளாக) குழுமப்பட வேண்டும் என்று இச்செயல்பாட்டால் தீர்மானிக்கப்படுகின்றது.

3. உறவைக் கண்டுபிடித்தல் (relation finding): ஒரு வாக்கியத்தில் உள்ள தொடர்பகுதிகள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனியலை, இடம் போன்றவை) தீர்மானிக்கப்படவேண்டும்.

2.5.3. தமிழில் சொல்வகை அடையாளப்படுத்திகள்

தமிழில் சொல் வகை அடையாளப்படுத்திகள் உருவாக்கப்பட்டுள்ளன. பின்வருவன அவற்றில் முக்கியமானவைகளாகும்:

1. வாசு அரங்கநாதனின் டேக் தமிழ்
2. கணேசனின் சொல்வகை அடையாளப்படுத்தி
3. RCILTS-தமிழின் கதம்பம்

2.5.4. தமிழில் தொடரியல் பகுத்துக் குறிப்பான்கள்

தமிழில் தொடரியல் பகுத்துக்குறிப்பான்கள் உருவாக்கும் முயற்சிகள் பல மேற்கொள்ளப்பட்டன. அவற்றில் சில இங்கு பட்டியலிடப்பட்டுள்ளன:

2.5.4.1. பாஸ்கரனின் தொடரியல் பகுத்துக்குறிப்பான்

முற்றுநிலைத் தானியங்கி மொழியமைப்பு மட்டத்தில் பகுத்துக்குறித்தலுக்கு முக்கியமான தொழில் நுட்பங்களில் ஒன்றாகும். பாஸ்கரன் (Baskaran, 1984) தமிழ் வாக்கியங்களில் பகுத்துக்குறித்தலுக்கு வேண்டி முற்றுநிலை தானியங்கையைப் பயன்படுத்தினார். இது தமிழ்ப்பல்கலைகழகத்தில் உருவாக்கப்பட்ட ருசியன்-தமிழ் இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு வேண்டி இவரால் உருவாக்கபட்டது.

2.5.4.2. குமார சண்முகத்தின் தமிழ்த்தொடரியல் பகுத்துக்குறிப்பான்

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக 'Parse Representaton of Tamil Syntax' என்ற தலைப்பில் ஒரு ஆய்வேட்டைச் சமர்பித்துள்ளார். இவ்வாய்வேட்டில் இவர் தமிழ்த் தொடரியல் பகுப்பாய்வு பற்றி வரிவான விளக்கம் தந்துள்ளார். சார்பு இலக்கண வடிவவாதம் அடிப்படையில் இவர் தமிழுக்கு ஒரு தொடரியல் பகுத்துக்குறின்பானையும் வடிவமைத்துள்ளார்.

2.6. தமிழில் மொழிநடை ஆய்வி

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர், சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி முலம் அறிய இயலும். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்பு பண்புக்கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த உரைகளை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்துக் கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனம் கண்டுக் கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய கணினி வழி தமிழ்மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியியல் மொழியியல் (Statistical linguistics), அளவீட்டு மொழியியல் (Quantitative linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணைப்புரிகின்றன. மொழி நடையை ஆய உதவும் மென்பொருள்கள் பல கண்டுப்பிடிக்கப்பட்டு பயன்பாட்டில் உள்ளன.

2.7. கணினிப் பொருண்மையியல்

பெருண்மையில் என்பது மொழிக் கூற்றுகளின் பொருண்மையை ஆய்வதாகும். நாம் இங்கு பொருண்மையைக் கண்டுபிடிக்க முறையான உருப்படுத்தங்களையும் கூற்றுக்களிலிருந்து இப்பொருண்மை உருப்படுத்தங்களுக்கு வேண்டி வழிமுறை வரைவுகளிலும் கவனக்குவிப்பு செய்யவேண்டும். நாம் எவ்வாறு கூற்றுக்களின் பொருண்மை தொடர்களின், சொற்களின் மற்றும் உருபன்களின் பொருண்மைகளுடன் தொடர்புடையவை என்று அறிந்திருக்கவேண்டும். பேசுபவர்கள் கேட்பவர்கள் மற்றும் கூற்றுக்கள் காணப்படும் சூழல்கள் இவற்றுடன் தொடர்புடைய சிக்கல்கள் பயன்வழியியலில் தீர்மானிக்கப்படும்.

2.7.1. பொருண்மையை உருப்படுத்தம் செய்தல்

மொழிக் கூற்றுக்களின் பொருண்மை முறையான அமைப்புகளில் வெளிப்படுத்தப்படும். இதை மொழிகளின் பொருண்மை உருப்படுத்தம் என்பர். இப்பொருண்மை உருப்படுத்தங்கள் ஒலியனியல், உருபனியல் மற்றும் தொடரியல் உருப்படுத்தங்களைப் போன்று பங்களிப்புச் செய்கின்றது.

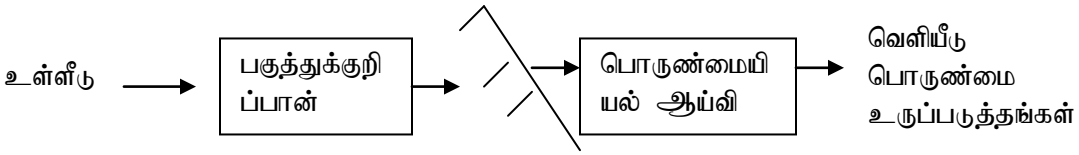
2.7.2. பொருண்மையியல் ஆய்வு

இங்கு பொருண்மை ஆய்வின் சிக்கல்களுக்கு கணினி அணுகுமுறை கருத்தில் கொள்ளப்படும். முதல் அணுகுமுறை தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வாகும் (Syntax-Driven Semantic Analysis). இதற்குக் குறைந்த அளவான பயன்பாடுதான் இருக்கின்றது. இது சொற்களஞ்சியத்திலிருந்தும் இலக்கணத்திலிருந்தும் நிலையான அறிவின் அடிப்படையில் உள்ளீடுகளுக்குப் பொருண்மை உருப்படுத்தங்களைத் தரும்.

2.7.3. தொடரியலால் இயக்கப்படும் பொருண்மையியல்

இது இணைப்பாக்கத்தின் கொள்கை அடிப்படையில் (principle of compostitionality) அமைந்தது. இவ்வணுகுமுறையில் உள்ளுறையும் முக்கிய கருத்து ஒரு வாக்கியத்தின் பொருண்மை அதன் பகுதிகளின் பொருண்மைகளால் இணைக்கப்பட்டது என்பதாகும். வாக்கியத்தின் பொருண்மை முழுவதும் அதில் அடங்கும் சொற்களின் அடிப்படையில் அமையாமல் வாக்கியத்தில் வரும் சொற்களின் நிரல், குழுமுதல் மற்றும் சொற்களுக்கிடையேயுள்ள தொடர்புகள் இவற்றின் அடிப்படையில் அமையும். அதாவது வாக்கியத்தின் பொருண்மை அதன் தொடரியல் அமைப்பின் அடிப்படையில் அமையும் என்பதாகும். எனவே தொடரியலால் இயக்கப்படும் பொருண்மையியல் ஆய்வில் பொருண்மை உருப்படுத்தங்களின் இணைப்பு ஒருவகை இலக்கணத்தால் தரப்படும் தொடரியல் கூறுகளாலும் உறவுகளாலும் வழிக்காட்டப்படும்.

உள்ளீடு செய்யப்படும் வாக்கியத்தின் தொடரியல் ஆய்வு பொருண்மையியல் ஆய்வின் உள்ளீடாக அமையும். கீழேத் தரப்பட்டுள்ளப்படம் இக்கருதலில் இருந்து நேரடியாகப் பெறப்படும் பைப்லைன் அடிப்படையிலான அணுகுமுறையை (pipeline-oriented approach) எடுத்துக்காட்டும்:



இவ்வாய்வு பொருண்மை ஆய்வின் உள்ளீடாகத் தரப்பட்டு பொருண்மை உருப்படுத்தத்தை உருவாக்கும்.

2.7.4. சூழல் கட்டுப்பாடில்லாத விதிகளில் பொருண்மையின் விரிவாக்கங்கள்

இக்கட்டுப்பாடில்லாத இலக்கண விதிகள் பொருண்மையியல் இணைப்புகளால் விரிவுப்படுத்தப்படும். இவ்விணைப்புகளை ஒரு கட்டுமானத்தின் உறுப்புகளின் பொருண்மையிலிருந்து அதன் பொருண்மை உருப்படுத்தம் கணிக்கப்படுகிறது என்பதைக் குறிப்பிடும் கற்றல் விதிகளாக கருத இயலும். விரிதாக்கப்பட்ட விதிகளுக்குப் பின்வரும் அமைப்பு இருக்கும்.

2.7.5. முந்தையப்பகுத்துக்குறிப்பானின் பொருண்மையியல் ஆய்வை ஒருங்கிணைத்தல்

முந்தைய பத்திகளில் நாம் பொருண்மையியல் ஆய்விகளின் எளிய பைப் லைன் கட்டமைப்பைப் பார்த்தோம். இதில் முழு தொடரியல் பகுத்துக்குறித்தலின் கிளைகள் பொருண்மையியல் ஆய்வுக்கு செலுத்தப்படுகிறது. இதன் நோக்கம் என்னவென்றால் இணைப்பாக்க அணுகுமுறை அது தொடர்புப்படுத்தும் தொடரியல் பகுத்துக்குறித்தலை வேண்டுகின்றது என்ற கருத்துச்சாயலிலிருந்து வந்ததாகும். இருப்பினும் பொருண்மையியல் ஆய்வை தொடரியல் ஆய்வுக்கு இணையாகச் செய்ய இயலும். முந்தையப் பகுத்துக்குறித்தலுடன் பொருண்மை ஆய்வின் இணைப்பு நேரிடையானதாகும். இணைப்பாக்கக் கொள்கை மரபுத் தொடர்களின் (idioms) பயன்பாட்டால் சிக்கலை நேரிடும். மரபுத்தொடர்கள் பிரித்து பொருண்மையை ஆய இயலாத மொழி அலகுகளாகும். இது இணைப்புக் கொள்கைக்கு நேரிடையான விளைவைத்தரும்.

2.7.6. சொற்பொருண்மையியல்

சொற்களஞ்சியம் (அகராதி) சொற்களின் பொருண்மை அமைப்பு மற்றும் அவைகளுக்கிடையே உள்ள உறவுகள் எவ்வாறு அமைந்துள்ளன என்பதை விளக்கும். இந்த ஒழுங்குமுறையான பொருண்மையியலில் தொடர்புடைய அமைப்பைப்பற்றி படிப்பது சொற்பொருண்மையியல் ஆகும். சொற்பொருண்மையியல் சொற்களுக்கிடையேயுள்ள உறவுகளை தொடர் உறவுகளாகவும் அடுக்குறவுகளாகவும் ஆயும். சொற்களுக்கிடையேயுள்ள அடுக்குறவுகள் ஒருபொருள்பன்மொழி, உள்ளடங்கு மொழி, உள்ளடக்கு மொழி, எதிர்மொழி போன்றவையாகும். மின்சொற்களஞ்சியங்களும் சொல்வலையும் சொற்களுக்கிடையேயுள்ள சொல்லுறவுகள் மற்றும் பொருண்மையுறவுகள் அடிப்படையில் அமைந்தவை.

2.8. தமிழில் சொற்களஞ்சியம் தயாரித்தல்

சொற்களஞ்சியம் (Thesaurus) தயாரிப்பதற்கு கணிப்பொறியைப் பயன்படுத்தலாம். ஒத்தல் (similarity), உள்ளடங்கல் (inclusion), மேலுறல் (overlapping), வேறுபடல் (disjunction) என்ற தர்க்கப்பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை அவற்றின் படிநிலை அமைப்பும் பொருண்மையும் வெளிப்படும் விதத்தில் பொருண்மைக் களங்களாகப் பகுத்துத் தருவது சொற்களஞ்சியம் ஆகும். கணிப்பொறி உதவியுடன் இச்சொற்களஞ்சியத்தை குறைந்த உழைப்பில் குறுகிய கால கட்டத்தில் தயாரித்து விடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்ததில் மின் சொற்களஞ்சியத்தை உருவாக்கலாம். இத்தகைய மின்சொற்களஞ்சியம் மொழிப்பெயர்த்தல், மொழிக் கற்பித்தல் போன்ற செயல்பாடுகளுக்கு பெரிதும் பயன்படும். தமிழுக்கான மின்சொற்களஞ்சியம் தமிழ்ப்பல்கலைக்கழக மொழியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து தயாரித்துள்ளன.

தரவு மையத்தில் சொற்கள் பொருண்மை அடிப்படையில் பகுப்பட்டு பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை வரிசைப்படுத்தித் தரவேண்டும். பின்வரும் எடுத்துக்காட்டு இதை விளக்கும்.

எ.கா.

1.பருமைப்பெயர்கள்

1.1 அ.ஃறிணைப்பெயர்கள்

1.1.1.உயிரில்லாதவை

1.1.1.1 இயற்கைப்பொருட்கள்

1.1.1.1.1விண்வெளிப்பொருட்கள்

1.1.1.1.1.1 வானம், ஆகாயம், வான், ஆகாசம், விண், விசும்பு, அந்தரிட்சம், அந்தரம்

1.1.1.1.1.2 கீழ்வானம், மேல்வானம்

இது தரவு மையமாக அமையும் இதிலிருந்து வழியமைப்புகள் வழிச் சொற்களஞ்சியத் தகவல்களைப் பெறலாம்.

2.9. தமிழில் சொல்வலை தயாரித்தல்

இது ஒரு இணையத் தளதரவு மையமாகும் (On line data base). இது சொற்களஞ்சியம் மற்றும் அகராதி என்ற இரண்டின் பண்புகளையும் கொண்டு மிகுந்த திறனுடையதாய் அமைந்துள்ளது. சொற்கள் தமக்கிடையே உள்ள பொருண்மை உறவுகளால் வலைபோல் பின்னப்பட்டிருப்பதை வெளிப்படுத்துவதால் இது அமையும். சொல் வலையின் மிகப் பரந்த நோக்கம் என்னவென்றால் சொல் தகவல்களை சொல் வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குப் படுத்துவதாகும். இதன் அடிப்படையில் சொல் வலை அகராதியிலிருந்து வேறுப்பட்டு சொற்களஞ்சியத்தையும் உள்ளடக்கும். ரொஜஸட் தெசாரஸ் போன் சொற்களஞ்சியங்களும், தமிழுக்கு இராஜேந்திரனால் (2001) உருவாக்கப்பட்ட தமிழ்ச்சொற்களஞ்சியமும் சொல் வலைப் பின்னுவதற்கு முன்னோடியாய் அமையும். இவையெல்லாம் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் அவற்றின் தேவைக்கு அதிகமான/மிகையான பதிவுகள் காணப்பெறும். இருசொற்கள், ஒருபொருள் பன்மொழிகள் என்றால் கூட அகரவரிக்காக இரண்டும் இருமுறைப் பதிவுப்பெறும். தலைப்பு ஒழுங்கு முறையில் அமைந்த சொற்களஞ்சியங்களின் சிக்கல் என்னவென்றால் இரண்டு முறை நோக்கீடு செய்து தான் தேவையான தகவல்களை பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எப்பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்து கொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவல்களைப் பெற இயலுகின்றது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். ஆனால் கணிப்பொறியின் தேடலை விரைவாக செய்ய இயலும். சொல் வலை மின்சொற்களஞ்சியத்தை விட ஒருபடி மேலானது. தமிழ்ப்பல்கலைக் கழகமும், சென்னையிலுள்ள AUKBC Research

Centrer-உம் இணைந்து தமிழ்சொல் வலையைத் தமிழ் இணையப் பல்கலைக்கழகத்திற்காக உருவாக்கியுள்ளது. இது முதல் கட்டநிலைதான். பொருளுதவி இல்லாத காரணத்தால் அடுத்தக் கட்டத்திற்குச் செல்ல இயலவில்லை.

2.10. தமிழில் சொற்களை அகரவரிசைப் படுத்துதல்

தமிழ்மொழியின் அகரவரிசைத் தொடர்பான செய்திகளையும் வழியமைப்பு ஆணைகளையும் (Programer Instructions) கணிப்பொறிக்குத் தந்தால் கணிப்பொறி மிகக் குறைந்த நேரத்தில் வேண்டிய சொற்களை அகரவரிசைப் படுத்தித் தந்துவிடும். ஆங்கிலத்திற்கு இது மிகச் சாதாரண ஒன்று. தமிழைப் பொருத்தவரையில் பலவிதமான எழுத்துருக்களும் (fonts) அதற்குப் பின்னணியாய் வேறுபட்ட குறியன்களும் (code) குறியனெண்களும் (code numbers) பயன்படுத்தப்படுவதால் அகரவரிசைப் படுத்தல் எளிய செயல்களாக அமைய வில்லை. இருப்பினும் வேறுப்பட்ட எழுத்துருக்களை வழியமைப்புகள் மூலம் இணைத்து அகரவரிசைப்படுத்தும் மென்பொருள்கள் கண்டுப்பிடிக்கப் பட்டு பயன்பாட்டில் உள்ளன.

2.11. தமிழில் சொல்லடைவு தயாரித்தல்

சொல்லடைவுத் தயாரித்தலையும் கணிப்பொறி மூலம் விரைவில் செய்துவிட இயலும். சொற்களையும் அவற்றோடு தொடர்பான செய்திகளையும் தந்தால் குறைந்த காலநேரத்தில் கணிப்பொறியின் உதவியால் சொல்லடைவைத் தயாரித்து விடலாம். எடுத்துக்காட்டாக, சங்க இலக்கியங்களுக்கு சொல்லடைவுத் தயாரிப்பதை எடுத்துக் கொள்வோம். பின்வரும் அடிகள் சங்க இலக்கியமான புறநானூற்றிலிருந்து எடுக்கப்பட்டது.

1. இவனியாரென்குவை யாயி னிவனே (புறம் 83.1)
2. தூங்குகையா னோங்கு நடைய (புறம் 22.1)

இவ்வடிகள் சொற்களாகப் பின்வருமாறுப் பிரித்துத் தரப்படவேண்டும். வழியமைப்புகள் மூலம் இச்சொற்களை அகரவரிசைப் படுத்தி அவை ஒவ்வொன்றும் வரும் இடங்களையும் குறிப்பிட்டுத் தர இயலும்.

ஆயின் (புறம் 83.1)

இவன் (புறம் 83.1)

என்குவை (புறம் 83.1)

ஏ (புறம் 83.1)

ஓங்கு (புறம் 22.1)

கையால் (புறம் 22.1)

தூங்கு (புறம் 22.1)

நடைய (புறம் 22.1)

ஒரு சொல்/சொல்லன் சங்க இலக்கியத்தில் எத்தனை முறை கையாளப்பெறுகின்றது. அல்லது குறிப்பிட்ட ஒரு நூலில் எத்தனை முறை கையாளப் பெறுகின்றது என்பன போன்ற தகவல்களை வழியமைப்புகள் மூலம் பெறலாம். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும், கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்தி சங்க இலக்கியங்களுக்கான ஒரு சொல்லடைவுத் தயாரித்துள்ளன.

2.12. தமிழில் சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவுத் தயாரித்தல்

சொல்லாட்சி அகராதி (concordance) ஒரு பனுவலில் இடம்பெறும் சொற்களை அகரவரிசைப் படுத்தித் தருவதோடு அச்சொற்கள் நூலில் இடம்பெறும் இடங்கள் எத்தனைமுறை இடம்பெறுகின்றன என்ற கணக்கீடு எந்தெந்த சூழலில் பயன்படுத்தப் படுகின்றன என்பன போன்ற செய்திகளைத் தரும். சொற்பொருளடைவு (compendium) சொல்லாட்சி அகராதிகளினின்றும் ஒரு படி மேற்சென்று சொற்கள் வரும் இடங்களையும், சூழல்நிலைகளையும் பொருள் அடிப்படையில் பிரித்துத் தரும் கணிப்பொறியைப் பொருத்தவரையில் இரண்டையும் ஒன்றில் அடக்கித் தந்து விடலாம். புறநானூற்றிலிருந்து எடுக்கப்பட்ட ஒரு சொல்லின் சொற்பொருளடைவின் மாதிரி கீழேத் தரப்பட்டுள்ளன.

அணி (Sense)

1 அழகு

1.1 அழகு தாழ்ந்த அணி(புறம் 57.1)

1.2 அணிகொள் தேர் (புறம் 68.14)

.....

.....

.....

2.அணிகலன்

2.1 இழையணிப்பொலிந்த வேந்து கோட்டல் குல்(புறம் 81.1)

.....

.....

3. ஒப்பனை

3.1 ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண் அணியல் (பதி 51.9)

4. அணிமை

5. வரிசை

6. திரட்சி

7. ஆடை

8. ஒப்புமை

9. சூழ்தல்

10. பக்கரை (சேணம்)

சொல்லடைவுத் தயாரிக்க முன்னர் கூறியதுபோல் செய்யுளின் அடியில் சொற்களாகப் பிரித்துத் தரப்படவேண்டும். இத்துடன் சூழல் அடையாளப்படுத்தப்பட வேண்டும். வழியமைப்புகள் மூலம் நாம் சொல்லாட்சி அகராதியையோ சொற்பொருளடைவையோ தயாரிக்க இயலும். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்தி சங்ககால இலக்கியங்களுக்கு இத்தகைய சொல்லாட்சி அகராதியும் சொற்பொருளடைவும் தயாரித்துள்ளன.

2.13. தமிழ்ச் சொல்லாய்வி தயாரித்தல்

தமிழுக்கானச் சொல் ஆய்விகள் பல வெளிவந்துள்ளன. அவை தமிழ்ப்உரைகளை உள்ளீடு செய்து வடிவமைத்து வெளியீடாகப்பெறும் நுட்பங்களைக் கொண்டுள்ளன. word, word star போன்ற மென்பொருள்களில் கிடைக்கும் எல்லா வசதிகளும் இவற்றிலும் உள்ளன.

சொற்களஞ்சியம், எழுத்து மற்றும் இலக்கணப் பிழைத்திருத்தி என்பன மேலும் மேம்படுத்தப்பட வேண்டும். பல தமிழ்ச்சொல் ஆய்விகளில் இவ்வசதிகள் இல்லை.

2.14. எழுத்துப் பெயர்ப்பு செய்தல்

ஒரு மொழியின் எழுத்துக்களால் எழுதப்பட்ட பிற மொழிகளின் எழுத்துக்களைப் பயன்படுத்தி படிக்கத்தக்க விதத்தில் மொழிப் பெயர்ப்புச் செய்யும் வசதிமுறைகள் கணிப்பொறியில் வந்து விட்டன. இதனால் தமிழ் எழுத்தில் எழுதப்பட்ட உரைகளை ரோமன் எழுத்தில் படிக்கவோ, ரோமன் எழுத்தில் எழுதப்பட்ட உரைகளை தமிழ் எழுத்தில் படிக்கவோ இயலும்.

2.15. தமிழுக்கு எழுத்துணரி உருவாக்குதல்

அச்சிடப்பட்ட உரைகளை அப்படியே கணினிக்கே எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியைப் பார்வை எழுத்தறிவான் (Optical charater recognitions) என்பர். இக்கருவி தட்டச்சு முறைபடி சாவிப்பலகை வழி உரைகளை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது. ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டு பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய கருவியை உருவாக்குவது கடினமான செயல். இருப்பினும் தமிழுக்கு இத்தகைய பார்வை எழுத்துணரி உருவாக்கும் முயற்சிகள் உருவாக்கப்பட்டு ஓரளவுக்கு வெற்றியும் அடைந்துள்ளன. கிருண்ஷ்ண மூர்த்தி அவர்கள் தமிழுக்கு மிகத் திறமையான பார்வை எழுத்துணரி உருவாக்கியுள்ளார்.

2.16. தமிழில் கணிப்பொறி வழி அகராதி தயாரித்தல்

மொழி கற்றலில் கணிப்பொறியின் பங்கு மகத்தானது. ஒரு மொழியின் சொல் அதன் இலக்கணக் கூறு, பொருள், பயன்பாடு மற்றும் வாக்கியத்தில் அதன் உபயோகம் போன்றவைகளை அகராதி விளக்குகிறது. மேலும் ஒரு சொல்லுக்கு நிகரான பிறமொழி பதங்களையும் இருமொழி அகராதிகள் விளங்குகின்றன. அகராதி தயாரிக்கும் பணி எளிதானதல்ல. மிகுந்த நேரமும், அதிக உடல் உழைப்பும் பொருட்செல்வம் ஆகும். ஒரு

அகராதியை உருவாக்க நீண்டகாலம் தேவைப்படும். எனவே அகராதி உருவாக்க கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன.

2.16.1. அகராதி உருவாக்கும் படிநிலைகள்

பொதுவாக அகராதி ஒன்றை தயாரிப்பதற்கு தேவைப்படும் படிநிலைகள்:

1. செய்திகளை சேகரித்தல், ஆராய்ந்து வகைப்படுத்துதல் அகர வரிசைப் படுத்துதல் என்பது முதற்கட்ட நிலைகளாகும்.
2. சொல்லுக்கான இலக்கணக்கூறு ஆராய்தல், பொருள் எழுதுதல், பிறமொழி சொற்களை எழுதுதல், சொல்லின் உபயோகத்தை விளக்குதல் இவைகளை இரண்டாம் நிலைப் பணியாகும்.

இதில் முதல் கட்ட பணி எளிமையாக கணிப்பொறி படுத்த இயலும். இரண்டாம் கட்டப்பணியை மனித முயற்ச்சியாலேயே செய்ய இயலும்.

2.16.2. கணிப்பொறி வழி அகராதி: வரலாற்றுச் சுருக்கம்

மாத்தூஸ் அர்னால் (Mathews Arnald) என்ற ஆங்கில கவிஞர் தன் கவிதைகளுக்கான சொல்லடைவை முதன் முதலாக கணிப்பொறி வழி உருவாக்கம் செய்தார். இதற்கு IBM-704 என்ற கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறி வழி சொல்லடைவினை முதன் முதலாக பேராசிரியர் ஸ்டிபன் (Stephen) உருவாக்கினார். இதனால் தொடர்ந்து Oxford அகராதி உருவாக்க சொல்லடைவு (Concordance) கணிப்பொறி வழி உருவாக்கப்பட்டது. இதைத் தொடர்ந்து ஆங்கிலம் மற்றும் பிற மொழிகளில் அகராதி உருவாக்க கணிப்பொறிகள் ஆங்காங்கே பயன்படுத்தப்பட்டன. தமிழ்ப்பல்கலைக்கழகத்தில் சங்க இலக்கியத்திற்கான' அகராதி உருவாக்க கணிப்பொறி பயன்படுத்தப்பட்டது. இதுவே முதல் முயற்சி ஆகும். கிரியா என்னும் தற்கால அகராதி கணிப்பொறி வழி உருவாக்கப்பட்டுள்ளது. இந்த அகராதி அச்சு வடிவிலும், குறுந்தகடுகளிலும் கிடைக்கப் பெறுகின்றன.

2.16.3. கணிப்பொறி வழி அகராதி: முறைமைகள்

கணிப்பொறி வழி அகராதி உருவாக்குவதற்கு கீழ்கண்ட முறைமைகள் பின்பற்றப்படுகின்றன.

1. செய்திகளை உள்ளீடு செய்தல்
2. செய்திகளை வார்த்தைகளாகப் பிரித்தல்

இந்நிலையில் ஒரு வாக்கியத்திலிருந்து சொற்கள் தனியாக பிரிக்கப்பட்டு பின்வரும் பட்டியல் உருவாக்கப்படுகிறது.

சொல்	வரி எண்
WORD	LINE NUMBER

1. மேற்கண்ட பட்டியலில் சொற்கள் அவைகள் இடம் பெறும் வரிசையில் இடம் பெறுகின்றன. மேலும் ஒரு சொல் உரையில் பல இடங்களில் பயனபடுத்தப்பட்டிருந்தால் அத்தனை முறை அச்சொல் இப்பட்டியலில் இடம்பெறும்.
2. மேற்கண்ட பட்டியல் அகரவரிசைப் படுத்தப்படுகிறது. இது மேற்கண்ட பட்டியலில் சொற்கள் இடம்பெறும் வரிசை முறையை மாற்றுகிறது. ஒரு சொல் பல முறை இடம் பெற்றிருப்பினும் அது அடுத்தடுத்து வரிசைமுறை பெறுகிறது.
3. இவ்வாறு அகரவரிசைப் படுத்தப் பட்டுள்ள பட்டுள்ள பட்டியலிலிருந்து கீழ்கண்ட பட்டியல் உருவாக்கப்படுகிறது.

சொல் (WORD)	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் இடம் LIST OF LINE NUMBER

இப்பட்டியலின் உதவிக் கொண்டு பின்வரும் பட்டியல் உருவாக்கப்படுகின்றது.

சொல் WORD	நிகழ்வெண்ணிக்கை FREQUENCY	சொல் இடம்பெறும் வரிசை எண் LIST OF LINE NUMBER	இடம் பெறும் வரி CITATION

மேற்கண்ட பட்டியலிலிருந்து ஒரு சொல் உரையில் எத்தனை முறை பயன்படுத்தப்பட்டுள்ளது என்பதை அறியலாம். அச்சொல் இடம்பெறும் வரி தெரிவதால் அதனைக்கொண்டு அச்சொல்லின் இலக்கணக்கூறு பொருள் உபயோகம் போன்றவைகளை ஆராய்ந்து எழுதலாம். இதன்மூலம் விவரங்கள் துல்லியமாகவும் விரைவாகவும் கிடைக்கப்பெறுகின்றன.

2.16.4. அகராதி தயாரித்தலில் இயற்கை மொழியாய்வு

கணிப்பொறி வழி அகராதி தயாரிக்கப்படும்போது கிடைக்கப்படும் இறுதி படிவம் சொல்லடைவு படிவம் போன்று இருக்கும். இதில் பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் மற்றும் வேற்றுமை உருபுகளுடன் கூடிய பெயர் சொற்கள் தனித்தனி சொற்களாகவே இடம்பெறும். உதாரணமாக கணிப்பொறி, கணிப்பொறியை, கணிப்பொறியில், கணிப்பொறியால், கணிப்பொறிக்கு போன்ற சொற்கள் தனிச்சொற்களாகவே இடம்பெறும். இவைகளை சொல் வடிவங்கள் என மொழியியலாளர்கள் கூறுகின்றனர். ஆனால் அகராதியில் இவைகள் தனிச்சொற்களாக இடம்பெறும் (அல்லது) இடம்பெறாது. கணிப்பொறி என்ற பெயர்ச்சொல் மட்டுமே அகராதியில் இடம்பெறும். இது போன்ற அகராதி சொற்கள் மொழியியலாளர்கள் LEXMES என்கின்றனர். இது போன்று படி என்ற வினைச் சொல்லுக்கு படித்தான், படித்தார், படித்தது போன்ற எண்ணற்ற சொல் வடிவங்கள் இருப்பினும் படி என்பதே அகராதிச் சொல்லாகும். கணிப்பொறி வழி சொல்லடைவு தயாரிக்கும்போது சொல் வடிவங்களே இடம்பெறும். இதிலிருந்து அகராதிச் சொற்களை மனித முயற்சியினாலேயே மேற்கொள்ள வேண்டும். இப்பணியையும் முழுமையாக கணிப்பொறிப்படுத்த உருபனியல் ஆய்வு மேற்கொள்ள வேண்டும். இவ்வாறாக சொல் வடிவத்திலிருந்து (Word Forms) உருபனியல் பகுப்பாய்வு மூலம் அகராதி சொற்களை LEXEMS DICTIONARY WORDS உருவாக்கும் முறையை மொழியியலாளர்கள் LEMMANTISATION என்று அழைக்கின்றனர். இம்முறையை கணிப்பொறிப்படுத்த முயற்சிகள் மேற்கொள்ளப்பட்டு வருகிறது. இதற்கு இயற்கை மொழி ஆய்வு உதவுகிறது.

2.16.5. முடிவுரை

கணிப்பொறி வழி அகராதி தயாரிக்கும் படிநிலைகளில் கணிப்பொறி படுத்த இயலும் நிலைகளையும், முறைகளையும் சுருக்கமாகப் பார்த்தோம். இம்முறைகளில் அகராதியில் மொழி

ஒரு தடையில்லை. தமிழ் அகராதியோ, ஆங்கில அகராதியோ, பிறமொழி அகராதிகளையோ மேற்கண்ட முறைகளைப் பின்பற்றி உருவாக்கலாம். அந்தந்த மொழிகளில் அகர வரிசைப்படுத்தப் படும் மென்பொருள்களை வடிவமைக்க வேண்டும்.

2.17. தமிழில் கணிப்பொறி வழி மொழி கற்றல், கற்பித்தல்

மக்கள் முன்னேற்றத்திற்கு கணிப்பொறி பல வகையலி உதவுகிறது. தொடக்கக் காலங்களில் கணிப்பொறி நல்ல கணக்கீட்டு கருவியாக (Calculating Device) பயன்பட்டது. அன்றாட பணிகளுக்கு அல்லது பணிகளிலும் வழக்கமாக செய்ய நேரிடும் பணிகளிலும் கணிப்பொறி பயன்படுத்தப்பட்டது. கணிப்பொறியின் வேகம், துல்லியம், செய்திகளை மீளத்தரும் வேகம் மற்றும் நினைவகத் திறன் போன்றவை அதிகரித்தன. ஆனால் அதற்கு எதிர் மாறாக விலை குறைந்தது. அதன் விளைவாகக் கணிப்பொறி பயன்பாடு அதிகரித்தது.

ஆக்கப் பூர்வமான (creative) பணிகளுக்கும் கணிப்பொறிகள் பயன்படும் வண்ணம் உருவாக்கப்படுகின்றன. குறிப்பாகக் கல்விக் கூடங்களில் பயிற்றுவிப்பதற்கு உதவும் கருவியாக கணிப்பொறிகள் பயன்படுத்தப்படுகின்றன. கணிப்பொறி வழி பாடங்களை தயாரித்து உரிய திட்டங்களை செயல்படுத்தி மாணவர்களுக்கு கற்பிக்கலாம். பாடங்களைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறி உதவுகிறது.

2.17.1. கற்றல், கற்பித்தலில் கணிப்பொறி

அடிப்படையில் கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. தேவையான பணிக்கான வழியமைப்பு அல்லது மென்பொருள் கொண்டு அது இயக்கப்படுகிறது. இருப்பினும் செய்திகளை சேகரித்தல், ஆராய்தல், வெளியிடுதல் மற்றும் பரப்புதல் போன்ற செயல்பாடுகளில் மிக மிக அதிவேகமுடையது. மேலும் கணிப்பொறியில் உரைச் செய்திகள், வரைபடங்கள் புகைப்படங்கள் ஒலிச்செய்திகள் மற்றும் சலனப்படங்கள் (animation) போன்ற பலவகையான செய்திகளை உள்ளீடு செய்து ஆராய்ந்து முடிவு பெறுவதற்கான வசதிகள் உள்ளன. எல்லா வகை செய்திகளையும் கலந்த கலவையாகவும் செய்திகள் அமையப்பெறலாம். இது போன்று எல்லா வகை செய்திகளையும கலந்த கலவையை பல்லூடகம் (multimedia) என்கிறோம்.

மற்றொரு பக்கத்தில் பயிற்றுவிப்போர் அல்லது ஆசிரியர்களுக்கு இருக்க வேண்டிய அடிப்படை இயல்புகள் மூன்று அவையாவன:

1. செய்திகளைத் திரட்டுதல்
2. செய்திகளை ஆராய்தல்
3. பயில்வோருக்கு ஏற்றவாறு செய்திகளை எடுத்துரைத்தல்

இத்தகைய குணங்களைக் கணிப்பொறி பெற்றிருப்பதால் கற்பித்தலுக்கும் கற்பதற்கும் சிறந்த சாதனமாகப் பயன்படுத்தப்படுகிறது. கணிப்பொறியால் சொந்தமாகப் பாடங்களை உருவாக்கவோ பாடங்களை நல்ல முறையில் வடிவமைப்பதற்கோ இயலாது. இவ்வாறான கணிப்பொறியை ஒரு ஆசிரியரைப் போன்று செயல்பட வைப்பதற்கு மென்பொருள் (soft ware) உருவாக்கப்படுகின்றன. ஒரு பாடத்தைப் பயிற்றுவிப்பதற்காக உருவாக்கப்படும் மென்பொருள் பொதுவாக பாட மென்பொருள் (course ware) எனப்படுகிறது.

கணிப்பொறி வழி கற்பித்தலுக்கு CAI (Computer Assisted/ Aided Instruction) என்று பெயர். இங்கு Assisted (or) Aided என்ற சொல் கற்பித்தலுக்கு கணிப்பொறி உதவும் என்ற கருத்தை வலியுறுத்தும். அதுபோல் கணிப்பொறி மூலம் பயிலும் முறைக்கு CAL (Computer Aided Learning) என்றுப் பெயர். CAL and CAI போன்ற சொற்களுக்கு நிகராக பல தொடர்கள் பயன்படுத்தப் படுகின்றன. CMI (Computer Managed Instruction), CML (Computer Managed Learning), CDL (Computer Directed Learning) என்பன சிலவாகும்.

2.17.2. கணிப்பொறி வழி கற்றல்-கற்பித்தலின் சிறப்பியல்புகள்

கணிப்பொறி வழி கற்றல், கற்பித்தலில் எண்ணற்ற பயன்களும், சிறப்பியல்புகளும் உள்ளன. பொதுவாக இச்சிறப்பியல்புகளை மூன்று வகைகளாகப் பிரிக்கலாம். அவையாவன,

1. கணிப்பொறியில் இயல்பாக இருக்கின்ற சிறப்பு குணங்கள்.
2. பயில்வோருக்கு கிடைக்கின்ற பயன்கள்.
3. பயிற்றுவிப்போருக்கு உதவும் சிறப்புகள்.

கணிப்பொறியின் செயல், வேகம், துல்லியம், சோர்வின்மை, சகிப்புத்தன்மை, நெகிழ்வுத் தன்மை, ஒரே நேரத்தில் பல செயல்களை புரியம் திறமை போன்றவை கணிப்பொறியில் இயல்பாக அமைந்துள்ள பண்பு நலன்களாகும். கணிப்பொறிகள் மனிதனைப் போல் சோர்வடைவதில்லை. எனவே எப்பொழுது வேண்டுமானாலும் எவ்வளவு நேரம் வேண்டுமென்றாலும் தொடர்நது பயன்படுத்தலாம். எனவே, கணிப்பொறிகள் கல்வி கற்றல் கற்பித்தலில் பயன்படுத்தப் பட்டு வருகின்றன. மேலும் வேறு எந்த ஒரு கருவியிலும் இல்லாத பல்லூடக சாதன அமைப்புக் கொண்டது. எனவே உரைநடைச் செய்திகள் மட்டுமல்லாது வரைப்படங்கள், புகைப்படங்கள் மற்றும் ஒலிச் செய்திகள் போன்றவைகளை ஆராய்வதற்கும், கையாளவதற்கும் எண்ணற்ற வசதிகளைப் பெற்றுள்ளது. இதனால் கல்வி கற்பிப்பதற்கும் கற்பதற்கும் தேவையான அனைத்து அடிப்படை வசதிகளையும் பெற்றுள்ளது.

பட வில்லைகள் (Slides) புகைப்படங்கள் மற்றம் கல்வித் தொடர்பான சினிமா படங்கள் கற்றலுக்கும், கற்பித்தலுக்கும் சிறந்த சாதனங்களாக பயன்படுத்தப்பட்டு வருகின்றன. இவைகளை கருவியாக பயன்படுத்திக் கொள்ள இயலுமெனினும் அந்த நிகழ்ச்சிகளுடன் உரையாடுவதற்கோ, கற்போரின் தன்மைக்கேற்ப கட்டுப்படுத்தவோ இயலாது. அந்நிகழ்ச்சிகளை தயாரித்த வரிசை கிரம முறையில் மட்டுமே பார்க்கலாம். கற்போரின் மனநிலைக்கு ஏற்ப தேவையான செய்திகளை கூடுதலாகவோ அல்லது குறைவாக பெறவோ இயலாது. ஆனால் கணிப்பொறி வழி இத்தகைய கல்வி நிகழ்ச்சிகள் அல்லது பாடங்களைப் பதிவு செய்யும் பொழுது அதனுடன் உரையாட இயலும். தேவையான செய்திகளை தேவையான வண்ணம் கேட்டுப் பெறலாம். இது கணிப்பொறி கருவியின் மிகப் பெரிய சிறப்பம்சமாகும்.

கணிப்பொறியை எந்நேரமும் எப்பொழுதும் பயன்படுத்தலாம் என்பதால் பயில்வோர் விருப்பமான நேரத்தில் விருப்பமான பாடத்தைப் பயிலலாம். கணிப்பொறி வழி பயில்வோரின் சுரந்திரம் தேவை. இங்கு பயில்வோர் தனித் தனியாக பயில வாய்ப்புகள் உண்டு. 1.பயில்வோரின் ஈடுபாடு, 2. கற்கும் திறன், 3. கற்கும் வேகம் ஆகிய தன்மைகளைப் பொறுத்து பாடங்களைக் கற்றுக் கொள்ளலாம். சாதாரண வகுப்புகளில் இத்தகைய சாத்தியங்கள் இல்லை. பாடங்களை விரைவாக கற்றுக் கெர்ளபவர்கள் மெதுவாக கற்கு கொள்பவரின் வேகத்திற்கு ஒத்துப்போக வேண்டும். இதனால் வேகமாக கற்றுக் கொள்பவரின் நேரம் வீணாகி போகும். வகுப்பில் கூச்ச சுபாவத்தால் ஐயப்பாடுகளை களைவதற்கான

கேள்விகளை கேட்பதற்கு மாணவர்கள் தயங்குவர். ஆனால் கணிப்பொறி வழி கற்கின்ற பொழுது மனநிலையைப் பொருத்து ஒரு பாடத்தில் அதிக செய்தியையோ, பொருளையோ, முக்கிய செய்தியையோ படித்து அறிவதற்கான பரிமாணங்களை அல்லது சிறப்புகளைப் பயில்வோரே நிர்மானித்துக் கொள்ளலாம். மேற்கூறியவைகள் எல்லாம் கணிப்பொறி வழி பயில்வோருக்கு உள்ள சிறப்புத் தன்மைகள் ஆகும்.

கணிப்பொறி வழி கற்பிப்பதற்கான பாடங்களை உருவாக்கும் ஆசிரியர்கள் மாணவர்களின் கல்வி நிலைக்கு ஏற்ப, கற்கும் திறனுக்கு ஏற்ப வடிவமைக்க வேண்டும் என்பதில்லை. ஒரு பொதுவான பாடங்களைத் தயாரித்து பயில்வோரின் திறமைக்கு ஏற்ப அவைகளை உருவாக்கித் தர இயலும். உரைநடைச் செய்திகள், விளக்கப் படங்கள் அனைத்தும் இதில் உள்ளதால் பாடக் கருத்துகளை எளிதாக புரிய வைக்க இயலும். ஒருமுறை பாடத் திட்டங்களை வகுத்து செம்மையாகத் தயாரித்து கணிப்பொறியில் எத்தனை முறை வேண்டுமானாலும் பயன்படுத்தலாம். ஆயிரக்கணக்கானவர்க்கு இத்தகைய பாடங்கள் பயன்படுத்துவதற்கு அமையும். இத்தகைய பாடங்களை ஒரு துறை நிபுணத்துவம் பெற்ற ஆசிரியர்கள் உருவாக்குகின்றனர். அவைகளில் திறமையானது ஆயிரக் கணக்கானோருக்கு உலகம் முழுவதும் பயன்படும்.

2.17.3. CAI- உருவாக்குவதற்கான கணிப்பொறி மொழிகள்

CAI – கணிப்பொறி வழி கற்பித்தல் மற்றும் கற்றல் பயன்பாடுகளுக்கான மென்பொருள்களை உருவாக்குவதற்கு என்றே கணிப்பொறி அமைப்பு மொழிகள் உள்ளன. IBM – (Internet Bussiness Machine) என்ற நிறுவனம் Course Writer என்ற வழியமைப்பு மொழியை இதுபோல் Control Data Corporation என்ற நிறுவனமும் TW என்ற மொழியை Starkwealther என்ற நிறுவனமும் Extol என்ற மொழியை Kenning and Kenning என்ற நிறுவனமும் தயாரித்து வெளியிட்டுள்ளன.

கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான பாடங்களை உருவாக்க மேற்கூரிய கணிப்பொறி மொழிகள் உதவுகின்றன. இவைகள் ஏனைய கணிப்பொறி மொழிகளை விட எளிமையானவை. இருப்பினும் ஒரு கணிப்பொறி மொழி கற்பதற்கான அதை பயன்படுத்துபவர்களுக்கான திறன் பெற்றவர்க்ள மட்டுமே இவைகளை பயன்படுத்த இயலும். எனவே இவைகளைக் கொண்டு பாடங்களை உருவாக்குபவர்கள் இக்கணிப்பொறி மொழியில்

தேர்ச்சி பெறுவது அவசியம் ஆகும். மேலும் வழியமைப்பு திறன் கொண்டவர்களாகவும் இருத்தல் அவசியம்.

2.17.4. CAI உருவாக்குவதற்கான பயனீட்டு மென்பொருள்கள் (packeges)

பொதுவாக வழியமைப்பு மொழிகளை பயன்படுத்தி ஒரு திட்டங்களை உருவாக்குகின்றபொழுது திட்டங்களை செயல்படுத்துவதற்கான முறைமைகளில் தேர்ச்சி பெற்றிருக்க வேண்டும். உதாரணமாக பத்து சொற்களை கணிப்பொறிக்கு தந்து அகரவரிசை பயன்படுத்துவதற்காக கொள்வோம். இப்பணியை செய்ய வழியமைப்பு மொழியில் தேர்ச்சி மட்டும் போதாது. அகரவரிசைப் படுத்துவதற்கான முறைகளை தெரிந்தவராகவும் இருத்தல் வேண்டும். அவ்வாறு இருந்தால் தான் அந்த முறைகளையும், செயல்பாடுகளையும் கணிப்பொறிக்கு கட்டளைகளாக செயல் படுத்த இயலும். ஆனால் பயனீட்டு மென்பொருள்களில் இதுபோன்ற செயல்பாடுகளுக்கு தேவையானவை தயார் நிலையில் உள்ளன. அவைகளைப் பயன்படுத்தி கொள்வதற்கான சாதாரண திறமைகள் இருந்தால் போதுமானது. எனவே, கணிப்பொறி வழி கற்றல் கற்பிப்பதற்கான மென்பொருள்கள் உருவாக்க கணிப்பொறி மொழிகளை பயன்படுத்த முனைவோர் பாடத் திட்டங்களை மட்டுமல்லாது கணிப்பொறி மொழியிலும் தேர்ச்சிப் பெற்றவராக இருத்தல் வேண்டும். இது ஒரு மிகப்பெரிய தடையாகும். இத்தடைகளை அகற்றவும், இப்பணிகளை எளிமைப் படுத்தவும் பல பயனீட்டு மென்பொருள் (Utility Soft Ware) வெளிவருகின்றன. இவைகளில் பாடங்களை கணிப்பொறி வழி தயார்படதற்கான அனைத்து வசதிகளும் தயார் நிலையில் உள்ளன. இத்தகைய பயனீட்டு மென்பொருள்களை எப்படி பயன்படுத்திக் கொள்வது என்ற பயிற்சி மட்டுமே போதுமானது. இவைகளை எளிதாகக் கற்றுக் கொள்ளலாம். கணிப்பொறி பயிற்சி இல்லாதவாகள் எளிமையாகவும், விரைவாகவும் கற்றுக் கொள்ள வசதி உள்ளது. Story Board மற்றும் Copy Write போன்றன CAI உருவாக்குவதற்கு பயன்படுத்தப்படும் பயனீட்டு மென்பொருள்கள் ஆகும். இவைகள் தவிர Light Source என்ற மென்பொருள்களும் உள்ளது.

2.17.5. CAI-இல் நகரும் படங்கள் (animation)

CAI மென்பொருள் உருவாக்கத்தில் நகரும் படங்கள் அல்லது சலனப்படங்கள் மற்றும் வீடியோ காட்சிகள் போன்றவைகளை அமைப்பது மென்பொருளின் தரத்தை மேம்படுத்தும். இதற்காக சில மென்பொருள்கள் உள்ளன. அவற்றுள் VIDEO WORKS என்பது ஒன்றாகும். இம்மென்பொருளில் சலனப்படங்களை உருவாக்கவும் கையாளவும் தேவையான வசதிகள் உள்ளன.

2.17.6. CAI உருவாக்கத்தில் மனித நுண்ணறிவின் பங்கு

கணிப்பொறி தானாக சிந்திக்கும் திறன் அற்றது. மனிதனைப் போன்று சிந்தித்து செயல்படக் கூடிய கணிப்பொறிகளை உருவாக்க அறிஞர்கள் முனைற்துள்ளனர். கணிப்பொறியை மனிதரைப் போல் உருவாக்கவும், சூழலுக்கு ஏற்ப பணியாற்றவும் அனுபவத்தில் புதிய செய்திகளை சுயமாக கற்கவும் தேவையான திறமைகளை உருவாக்கி செயல்படுமாறு உருவாக்குகின்றன. இத்துறைக்கு செயற்கை நுண்ணறிவு (artificial intelligence) என்று பெயர். இதை AI என்று சுருக்கமாக அழைப்பர். இத்துறையின் ஒரு அங்கமாக இயற்கை மொழியாய்வு NLP (Natural Language Processing) என்ற துறை உள்ளது. ஒரு மனிதன் மற்றவர்களிடம் உரையாடுவது போல் மனித மொழியில் கணிப்பொறியுடன் உரையாட தேவையான வசதிகளை இத்துறை செய்கிறது. இந்த ஆய்வு வெற்றி பெறுமேயானால் கணிப்பொறிக்கு கட்டளையிட கணிப்பொறி மொழி என்ற ஒன்று தேவையில்லை. மனித மொழியில் கணிப்பொறிக்கு கட்டளை தரலாம். அதை கணிப்பொறி புரிந்துக் கொண்டு செயலாற்றும். அதுபோல் கணிப்பொறியின் விடையை மனித மொழியில் விளக்கமாகத் தரும். இத்தகைய அமைப்புகள் மனிதனுக்கும், கணிப்பொறிக்கும் இருக்கின்ற உறவை எளிமைப்படுத்தும். மேலும் துறை ஒருங்கு (Expert System) ES என்பது கணிப்பொறி எழுதியுள்ள மிகப்பெரிய பிரிவாகும். கணிப்பொறியை ஒரு துறையின் வல்லுநரைப் போன்று பயன்படுத்த இவ்வாய்வுகள் உதவும்.

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருள்களை உருவாக்கும் பொழுது மேற்கூறிய மொழி நுட்பங்கள் மனித நுண்ணறிவு இயற்கை மொழி ஆய்வு துறை ஒருங்கு போன்றன சேர்க்கப்படுகின்றன. செயற்கை நுண்ணறிவுடன் கூடிய CAI மென்பொருளை ACAI என்று அழைக்கிறார்கள். இதை ITS (Intelligence Tutorial System) என்றும் அழைப்பர்.

2.17.7.கணிப்பொறி வழி மொழி கற்றல், கற்பித்தலின் எல்லைகள்

கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் கணிப்பொறி சிறந்து பங்காற்றுகிறது. ஆனால் மொழி கற்றல் கற்பித்தல் போன்ற செயல்களில் எண்ணற்ற நுணுக்கங்கள் தேவைப்படுகின்றன. இதில் கணிப்பொறி ஓரளவுதான் பங்காற்ற முடியுமே தவிர பயிற்று விக்கும் ஆசிரியருக்கு நிகராக அமையும் அல்லது அமையாது. மொழி கற்பிப்பதற்கு கணிப்பொறி துணை கருவியாக மட்டும் பயன்படும். இருப்பினும் இத்தகைய மென்பொருள் உருவாக்குகின்ற பொழுது எது இயலுமோ அதில் உறுதியும், திடமான சிந்தனையும் வேண்டும். மொழி கற்றல் கற்பித்தலில் கணிப்பொறியில் செய்யத் தரும் பயிற்சிகள் கீழேப் பட்டியலிடப் பட்டுள்ளன:

- கணிப்பொறி வழி உரையாடல் பயிற்சி அளித்தல்.
- மொழியில் கோடியிட்ட இடங்களை நிரப்புதல் போன்ற வினாத்தாள்களின் பயிற்சி.
- ஒரு சொல்லைத் தந்து அதற்கான உச்சரிப்பு.
- வாக்கிய வகைளை மாற்றி அமைத்தல்.
- இரு சிறு வாக்கியங்களை இணைத்து ஒரு வாக்கியமாக உருவாக்குதல்.
- சொல்லின் ஒருமை, பன்மை, எதிர்பதங்கள் போன்றவற்றை கற்று கொள்வதற்கான பயிற்சிகள்.
- சொற்களை பிரித்து எழுதுதல்
- சொற்களை சேர்த்து எழுதுதல்
- ஒரு சொல்லின் இலக்கணக் கூறினைக் கொடுத்து மற்ற இலக்கணக் கூறுகளைக் கேட்டல்.

மேற்கூறிய பட்டியலில் கணிப்பொறி வழி கற்றல் கற்பித்தலுக்கான சில வாய்ப்புகள் கொடுக்கப்பட்டுள்ளன. இதுபோன்று எளிமையான பயிற்சிகளை தருவதற்கு மென்பொருள்களை உருவாக்கி பயன்படுத்தலாம்.

2.17.8.கணிப்பொறி வழி தமிழ் கற்றல்-கற்பித்தலுக்கான மென்பொருள்கள்

கணிப்பொறி வழி கறிபத்தலுக்கு எண்ணற்ற மென்பொருள்கள் உள்ளன. அறிவியல் தொழில் நுட்பம் போன்ற பாடங்களை பயிற்றுவிப்பதற்கு உலகளவில் பல்லாயிரக்க கணக்கான மென்பொருள்கள் உருவாக்கப்பட்டு பயன்படுத்தப்பட்டு வருகின்றன. அதுபோல் மொழி கற்றல்

கற்பித்தலுக்குப் பல மென்பொருள்கள் உள்ளன. கணிப்பொறி வழி தமிழ் கற்றல் கற்பித்தலுக்கு மெனபொருள் உருவாக்கம் தொடங்கப்பட்டுள்ளது.

சென்னையைச் சார்ந்த Learn Fun என்ற நிறுவனம் தமிழுக்காக மென்பொருள் ஒன்றை உருவாக்கி உள்ளது. இம்மென்பொருளில் தமிழ் எழுத்துக்களை எழுத, பழக வசதிகள் உள்ளன. இம்மென்பொருள் மொழி விளையாட்டு மென்பொருள் வகையைச் சேர்ந்தது. தமிழை ஆரம்ப நிலையில் கற்பவர்கள் வல்லினம், மெல்லினம், குறில், நெடில் போன்ற அடிப்படைக் கூறுகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் மிகவும் உதவுகிறது.

சென்னையைச் சார்ந்த Soft View என்ற நிறுவனம் அமுதம் என்ற மெனபொருளை உருவாக்கியுள்ளது. இம்மென்பொருள், திருக்குறள், ஆத்திச் சூடி போன்றவைகளை எளிமையாகக் கற்று தருகிறது. சிங்கப்பூர் தேசிய கல்விக் கழகம் தமிழில் பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை உருவாக்கியுள்ளது. ஒருமை, பன்மை, எதிர்ப்பதம் போன்றவைகளை கற்றுக் கொள்வதற்கு இம்மென்பொருள் உதவுகிறது.

சிங்கப்பூரைச் சார்ந்த தமிழாசிரியர் கழகம் தமிழைப் பயிற்றுவிப்பதற்காக சில மென்பொருளை வெளியிட்டுள்ளன. அவைகளைக் கொண்டு தமிழில் உள்ள இரண்டு எழுத்துச் சொற்கள், மூன்று எழுத்துச் சொற்கள் போன்றவைகளை எளிமையாக கற்கலாம். மேலும் வெளிப்படும் படத்திலிருந்து அப்படத்தில் உள்ள பொருளை அடையாளம் காண்டு பயிற்சிப் பெற சில வசதிகள் இம்மென்பொருளில் உள்ளன. மேலும் ஒன்றியிலிருந்து ஆயிரம் வரை எண்கள் எழுதவும், படிக்கவும் சொல்லித் தருகிறது. மேற்கண்ட மென்பொருள் தயாரிக்கும் பணி அமரர் பேராசிரியார் கோவிந்தசாமி தலைமையில் சிங்கப்பூரில் நடைப்பெற்றது.

அமெரிக்காவைச் சார்நத லிண்ட்ஹோம் என்பவர் தமிழ்ச் சொற்கோவைகளை பயிற்றுவிப்பதற்காக மென்பொருள் ஒன்றை வெளியிட்டுள்ளார். பென்சில்லேனியா பல்கலைக் கழகத்தில் பேச்சுத் தமிழைப் பயிற்றுவிப்பதற்காக மென்பொருள் உருவாக்க முனைந்துள்ளனர். அமெரிக்காவில் உள்ள சிக்காக்கோ நகரிலுள்ள நிறுவனம் ஒன்று Visual Tamil என்ற மென்பொருளை உருவாக்கியுள்ளது. இம்மென்பொருள் தமிழ் கலை, பண்பாடு ஆகியவற்றைக் கற்பிக்கிறது. இது போன்று ஆங்காங்கே கணிப்பொறி வழி தமிழில் கற்பித்தல் கற்றலுக்கு பல மென்பொருள்கள் உருவாக்கப்படுகின்றன.

2.17.9. கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலில் இணையத்தின் பங்கு

ஒன்றுக்கு மேற்பட்ட கணிப்பொறிகள் ஒன்றோடு ஒன்று இணைக்கப்படுகின்றன. இத்தகைய அமைப்பிற்கு கணிப்பொறி வலைப்பின்னல் (Computer Net work) என்று பெயர். இதனால் ஒரு கணிப்பொறியில் பதிவு செய்யப் பட்டுள்ள மென்பொருள் மற்றும் பதிவுகளை கணிப்பொறிகளில் பெறலாம். இது போன்று உலகளாவிய நிலையில் உள்ள கணிப்பொறிகளை ஒன்றிணைத்து தகவல் பரிமாற்றம் செய்து கொள்ள உதவும் அமைப்பிற்கு இணையம் என்று பெயர். இதன் பயனாய் இணையத்தில் உள்ள ஒரு கணிப்பொறியில் பதிவு செய்யப்படும் செய்திகளை உலகில் உள்ள எந்த கணிப்பொறியிலிருந்தும் பெறலாம். கணிப்பொறி வழி மொழி கற்றல் கற்பித்தலுக்கான மென்பொருள் இணையம் ஏற்றம் செய்யப்பட்டுள்ளன. இம்மென்பொருள்களை செயல்படுத்தி மொழிகளை கற்பிக்கலாம். பென்சில்லேனியா பல்கலைக் கழகத்தில் WALTT – Web Aided Language Teaching and Learning அல்லது Training என்ற திட்டம் செயல்படுத்தப் பட்டுள்ளது.

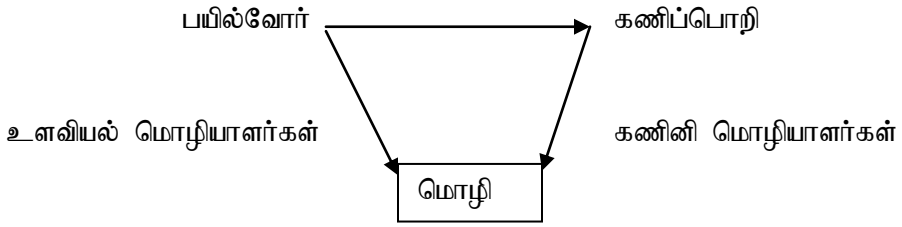
இணையத்தின் வழி பேச்சு தமிழை பயிற்றுவிப்பதற்கு இது உதவுகிறது. மேலும் தமிழ் மொழியை கற்பிப்பதற்கு சிங்கப்பூரில் உருவாக்கப்பட்டுள்ள பல இணையத் தளங்கள் செம்மையாக உள்ளன. தமிழ்ப்பல்கலைக் கழகத்தின் அங்கமாக தமிழ் இணையப் பல்கலைக்கழகம் செயல்படுகிறது. இணையப் பல்கலைக்கழகம் தமிழ் பாடத்தில் சான்றிதழ் வகுப்புகளையும், பட்டய வகுப்புகளையும், தொலைத்தூர கல்வி திட்டத்தின் கீழ் இணையம் வழி நடத்துகிறது. அயல் நாட்டில் உள்ள தமிழர்கள் மற்றும் இந்திய மாநிலங்களில் உள்ள தமிழர்கள் தமிழ் மொழியை இதன் வழி கற்கலாம். இப்பல்கலைக் கழகம் தமிழ் மொழி கலாச்சாரம், பண்பாடு போன்றவற்றை இணையம் வழி கற்பிக்கிறது. மேலும் தமிழ் மொழியை கற்றுக் கொள்வதற்கு உறுதுணையாக அமையும் அகராதிகள், இலக்கியங்கள் போன்றவைகளை இந்த இணையத்தின் மூலம் இலவசமாக கற்கலாம். இது போன்று தமிழ்மொழி கற்றலுக்கும், கற்பித்தலுக்கும் இணையத்தில் பல வாய்ப்புகள் உள்ளன.

2.17.10. மொழி கற்றல் மென்பொருள் உருவாக்கத்தில் பிற துறையினரின் பங்கு

இத்தகைய மென்பொருளை உருவாக்க மொழி வல்லுநர் உளவியல் மொழியியாளர்கள் (Psychology Linguistics) கணினி மொழியியல் வல்லுநர்கள் (Computational Linguistics)

ஆகியோரின் ஒத்துழைப்பு அவசியமாகும். மொழியியல் வல்லுநர்கள் கணிப்பொறிக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். உளவியல் மொழியாளர்கள் பயில்வோருக்கும், மொழிக்கும் உள்ள உறவை ஆராய்கின்றனர். மேலும் கணிப்பொறிக்கும், பயில்வோருக்கும் உள்ள உறவை ERGNOMICS என்ற துறையினர் ஆராய்கின்றனர். இத்தொடர்பு கீழே கண்ட கட்டப்படம் விளக்குகிறது.

Ergnomics



கணிப்பொறி வழி கற்றல், கற்பித்தலில் மேற்கூறியவாறு கணிப்பொறி பல வகைகளில் பயன்படுகிறது.

2.18. கணிப்பொறி வழி மொழிபெயர்ப்பு

பாரதத்தில் பன்மொழி பேசும் மக்கள் உள்ளன. மொழி, இனம், பண்பாடு மற்றும் கலாச்சாரத்தால் பல்வேறு மக்களின் எண்ணங்கள், கருத்துக்கள் பிறமொழி பேசும் மக்களிடையே சென்றடைவது அவசியமாகும். அதுபோல் ஒருமொழி பேசும் மக்களின் பண்பாடு, கலாச்சாரம், இலக்கியம் போன்றவைகளாகும். பிறமொழி பேசும் மக்களிடையே சென்றடைய வேண்டும். அப்பொழுதுதான் மக்களிடையே ஒற்றுமை உணர்ச்சியும் தேசிய ஒருமைப்பாடும் நிகழும். ஒரு மொழியில் உள்ள செய்திகள் பிற மொழிகளில் மாற்றம் செய்யப்படுவது மிகவும் அவசியமாகும். இதுபோல் உலக நாடுகளில் ஒரு நாடு பிற நாடுகளிலிருந்து தனித்து இருக்க இயலாது. தொழில் நுட்ப மேம்பாடு, பொருளாதார மேம்பாடு, சமூக மேம்பாடு போன்றவைகளுக்கு ஒன்றோடு ஒன்று தொடர்பு கொண்டு கூட்டு முயற்சியில் மனித வாழ்வியலில் மேம்பாட்டினைக் கருத்தில் கொள்ள வேண்டும்.

ஒரு நாட்டின் கலாச்சாரம் பண்பாடு போன்ற பிற நாடுகளுக்கும் பரவ செய்ய வேண்டும். ஒரு நாட்டின் இலக்கியச் செல்வங்கள் பிற நாடுகளைச் சென்றடைய வேண்டும்.

அப்பொழுதுதான் ஒரு நாட்டின் தொன்மை, தனித்துவம், வளம் போன்றவைகள் பிற நாடுகளுக்கு தெரிய வரும். எனவே ஒரு நாட்டில் உருவாக்கப்படும் படைப்புகள் பிற நாடுகளுக்கு எளிதில் சென்றடைய வேண்டுமெனில் மொழிபெயர்ப்புப் பணி அவசியமாகும். நாம் தகவல் யுகத்தில் வாழ்ந்து வருகிறோம். செய்திகள் மலையெனக் குவிகின்றன. இச்செய்திகளை உடனுக்குடன் மொழிபெயர்க்க வேண்டியது அவசியமாகிறது. வாழ்வில் எல்லா நிலைகளலும் பயன் தரக் கூடிய கணிப்பொறி மொழி பெயர்ப்புக்குப் பயன்படுத்த மேற்கொள்ளப் பட்டுள்ளன.

ஒரு மொழியில் உள்ள செய்திகளை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பு இயல்பு தன்மை மாறாது இருத்தல் வேண்டும். செய்தி எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Languages) என்கிறோம். செய்திகள் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியை இலக்கு மொழி என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியில் மேற்கொள்ளுவதென்பது மிகக் கடினமானப் பணியாகும். எனவே மொழி பெயர்ப்புகளை கணிப்பொறி வழி மொழிபெயர்ப்பு (Machine Translation) மேற்கொள்வதற்கான தேவைகள் மற்றும் அவசியங்கள் கீழே பட்டடியலிடப் பட்டுள்ளன:

- உலக அளவில் செய்திகள் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழி பெயர்த்தல் அவசியமாகிறது.
- கணிப்பொறி வழி மொழி பெயர்த்தல் மனித உழைப்பை குறைக்கிறது.
- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்பு பணிகளை செய்ய இயல்கிறது. இதனால் கால விரயம் தவிர்க்கப் படுகிறது.
- தற்போதைய மொழி பெயர்ப்பு பணிகளுக்கு ஏற்ற வகையில் மொழிபெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மூலமொழி பெயர்க்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழி மற்றும் இலக்கு மொழிகளில் புலமை வாய்ந்தவராக இருத்தல் வேண்டும். ஆனால் கணிப்பொறியில் இந்தக் கட்டுபாடு இல்லை. தேவையான மொழியில் புலமை உள்ளவர்களிடம் கருத்துகள் பெற்று அதை வழியமைப்பு மொழிகளில் எழுதி மொழி பெயர்ப்பு செய்யலாம்.

2.18.1. கணிப்பொறி மொழி பெயர்ப்பின் வகைகள்

பொதுவாக கணிப்பொறி வழி மேற்கொள்ளப்படும் மொழிபெயர்ப்பு பின் வரும் முன்று வகைகளாகப் பிரிக்கப்படுகின்றன:

1. கணிப்பொறி உதவியுடன் கூடிய மனித மொழி பெயர்ப்பு (CAHT – Computer Assisted Human Translation)
2. மனித உதவியுடன் கூடிய இயந்திர மொழி பெயர்ப்பு (HAMT – Human Assisted Machine Translation)
3. முற்றிலும் தானியங்கி (அல்லது) இயந்திர மொழிபெயர்ப்பு (AMTS – Automatic Machine Translation System)

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனித மொழி பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகிறது. அவர்கள் மொழி பெயர்ப்பு பணியில் ஈடுபடும் போது ஒரு மொழியில் உள்ள சொற்களுக்கு நிகரான பிறமொழிச் சொற்களை தெரிந்து கொள்ளவும் அச்சொற்களின் பொருள், பயன்பாடு ஆகியவற்றை எளிமையாக தெரிந்து கொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாக தயார் நிலையில் தெரிந்து கொள்வதற்கு கணிப்பொறி உதவுகிறது. இவ்வகை மொழிப்பெயர்ப்புகளில் மொழிப்பெயர்ப்பாளர்கள் முதல் நிலையிலும், அவர்களுக்கு உதவக் கூடிய வகையில் கணிப்பொறி பயன்படுகிறது.

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிபெயர்ப்பு பணியை செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ளன. தகவலுக்கு மேலாக ஏதாவது செய்தறிகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட முவகை மொழிபெயர்ப்பு வகைகளில் மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பே நடைமுறையில் உள்ளது.

முற்றிலும் இயந்திர மொழி பெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல. இதற்கு எண்ணற்ற மொழியின் கோட்பாடுகளும், மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட வேண்டியது கற்று இருப்பினும் கணிப்பொறி வழி மொழி பெயர்ப்பில் முற்றிலும் தாணியங்கப் படுத்தும் முயற்சி இலக்காக கொள்ளப்பட்டுள்ளது.

2.18.2. கணிப்பொறி வழி மொழி பெயர்ப்பின் வரலாறு

கணிப்பொறி வழி மொழபெயர்ப்பினை தானியங்கி மொழிபெயர்ப்பு என்றும், இயந்திர மொழி பெயர்ப்பு என்றும், கணிப்பொறி உதவியுடன் நடைபெறும் மொழி பெயர்ப்பு என்றும் கொள்ளலாம். கணிப்பொறியின் வேகம், நினைவகத் திறன், பதிவு செய்துள்ள செய்திகளை தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளபெறக் கூடிய வசதி போன்றன கணிப்பொறி வழி மொழி பெயர்ப்பு பணியை மேற்கொள்ள உதவின.

கணிப்பொறி வழி மொழி பெயர்ப்பு பணி 1950 –ம் ஆண்டு வாரன் வீவர் (Warren Weiver) என்ற அறிஞரால் தொடங்கப்பெற்றது. இதைத் தொடர்ந்து ஓ பார் ஹில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்பு பணியை ஆரம்பித்தார். 1954-ம் ஆண்டு ஜார்ஜ் டவுன் பல்கலைக் கழகம் கணிப்பொறி வழி மொழி பெயர்ப்பு பணியை துவங்கியது. இதைத் தொடர்ந்து 1956-ம் ஆண்டு மாஸ்கோவில் பிரெஞ்சு மொழிகளிலிருந்து ரஷ்ய மொழிக்கு கணிப்பொறி வழி மொழிபெயர்ப்பு பணி மேற்கொள்ளப்பட்டது. இப்பணியை மேல் ஜங்க் (Melchunk) என்ற அறிஞர் மேற்கொண்டார்.

1973-ம் ஆண்டு அமெரிக்காவில் உள்ள LOGOS Incorporation Limited நிறுவனம் மொழி பெயர்ப்பிற்கான மென்பொருளை வடிவமைத்தது. இம்பென்பொருளைக் கொண்டு ஆங்கிலத்தில் அமைந்துள்ள செய்திகளை வியட்நாம் மொழியில் மொழி பெயர்ப்பு செய்யப்பட்டது. ஏறக்குறைய இதே காலக் கட்டத்தில் Tertils என்ற நிறுவனம் TITUS என்ற மென்பொருளை வடிவமைத்தது. இம்மென்பொருளைக் கொண்டு ஆங்கிலம், ஸ்பானிஷ், பிரெஞ்ச் மற்றும் ஜெர்மனி ஆகிய மொழியில் உள்ள செய்திகள் ஒனிறிலிருந்து மற்றவைகளுக்கு மொழி மாற்றம் செய்யப்பட்டது. இதைத் தொடர்ந்து CUTD CHINESE UNIVERSITY TRANSLATOR என்ற மென்பொருள் உருவாக்கப்பட்டது. இதைக் கொண்டு சீன மொழியில் உள்ள செய்தித் தாள்கள் ஆங்கில மொழிக்கு மொழி மாற்றம் செய்யப்பட்டன.

1974-ம் ஆண்டு SYSTRAN என்ற மென்பொருள் உருவாக்கப்பட்டது. SYSTRAN என்பதன் விரிவு System Translation என்பதாகும். இம்பொழி பெயர்ப்பு மென்பொருள் உருஷ்ய மொழியிலிருந்து பிரெஞ்ச் மொழிக்கு மொழி மாற்றம் செய்யப்பட்டது. இதற்கானத் திட்டத்தை

P. TOMA என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன் முதலில் வெளியிடப்பட்ட கணிப்பொறி வழி மொழி பெயர்ப்பிற்கான மென்பொருள் இதுவேயாகும்.

1975-ம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென்பொருள் உருஷ்ய மொழியில் அமைந்துள்ள செய்திகளை பிரெஞ்ச் மொழிக்கு மாற்றம் செய்யப்படுவது. இதே காலக் கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு மொழிபெயர்க்க வடிவமைக்கப்பட்டது. இதுபோன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனை கட்ட முயற்சியாகவும், வணிக நோக்கத்திலும் கணிப்பொறி வழி மொழி பெயர்ப்பிற்கான மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைப்பெற்று வருகின்றன. 1985-ம் ஆண்டு தமிழ்பபல்கலைக் கழகத்தில் TUMTS (Tamil University Translation System) என்ற திட்டப்பணி தொடங்கப் பெற்றது. உருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிபெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக் கழகங்களில் கணிப்பொறி அறிவியல் துறை, மொழியில் துறை மற்றும் மொழிபெயர்ப்பு துறை ஆகியவைகளின் கூட்டு முயற்சியில் விளைந்ததாகும்.

இதற்குப் பின்னால் கான்பூரில் உள்ள இந்திய தொழில் நுட்ட நிறுவனம் (Indian Institute of Technology) ஹிந்தியிலிருந்து ஆங்கிலம், ஆங்கிலத்திலிருந்து ஹிந்தியை மொழி பெயர்க்கவும் இந்தியிலிருந்து(தமிழ், தெலுங்கு, கன்னடா போன்ற) இந்திய மொழிகளுக்கு திட்டமிட்டு செயலாற்றியது. இப்போது ஹைதராத்தில் உள்ள உலகத் தகல் தொழில் நுட்பநிறுவனம் (Indian Institute of Information Technology (IIIT), பத் முழுமுயற்சியுடன் இயந்திர மொழி பெயர்ப்பில் ஈடுபட்டுள்ளது. தற்பொழுது சென்னையிலுள்ள AU- Anna University , KRC- Kalyana Sundharam Reasch Center என்ற ஆய்வு நிறுவனம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழி பெயர்க்க திட்டம் ஒன்றினை மேற்கொண்டு செயல்படுத்தி வருகிறது.

சமீபத்தில் கருத்துப்பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் தகல் தொழில் நுட்ப நிறுவனத்தின் நிதி நல்கையில் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கும் மொழிபெயர்ப்பு செய்ய வேண்டி இரண்டு குழுக்கள் (consortium) அமைக்கப்பட்டு செயல்பட உள்ளன. இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு மொழிபெயர்க்கும் குழுவின் தலைவராக பேராசிரியர் இராஜிவ் சங்கல் அவர்கள் செயலாற்றுகின்றார். இக்குழுவின் அங்கத்தினாகளின் ஒருவராக தமிழ்ப்பல்கலைக்கழக

மொழியில் துறை பேராசிரியர் ச.இராசேந்திரன் அவர்கள் இருக்கின்றார். இவரிடம் மலையளம்-தமிழ் மொழிபெயர்ப்பு ஒப்படைக்கப்பட்டுள்ளது.

2.18.3. கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிகளுக்கு, மொழிமாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்தகைய கணிப்பொறி வழி மொழிபெயர்ப்பு பணிகளில் பின்வரும் இரு எல்லைகள் நிர்ணயிக்கப் படுகின்றன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் வட்டார அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக் கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்துவது என்பதற்கோ, அதே பொருளில் பயன்படுத்துவது என்பதற்கோ உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழியை இலக்கிய மொழி, பேச்சு மொழி என இரு வகையாகப் பிரிக்கப்படுகிறது. பேச்சு மொழியை கணிப்பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிபெயர்ப்பு பணிகளில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக் கொள்ளப்படுகின்றன.

2. செய்யுள் கவிதை போன்றவைகளில் பொருளுக்கு தரும் முக்கியத்துவத்தை விட கருத்தினைப் புலப்படுத்தும் சக்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகினறன. இவைகள் மனித உளவியலை பிரதிபலிப்பான இவ் உளவியல் கூறுகளை கணிப்பொறி வழி வெளிப்படுத்துவதென்பது இயலாத ஒன்றாகும். இதுபோன்று இலக்கியங்களிலும் பொருளை விட உளவியல் கூறுகளுக்கு முக்கத்துவங்கள் தரப்படுகின்றன. எனவே செய்யுள், கவிதை, பாடல்கள் மற்றும் இலக்கிய செய்திகள் ஆகியவைகளை கணிப்பொறிற வழி மொழிப்பெயர்ப்பிற்கு தொடக்க நிலையில் எடுத்துக் கொள்ளப்பட வில்லை. அறிவியல் தொழில் நுட்பம் சார்ந்த படைப்புகளில் பொருளுக்கே முக்கியத்துவம் தரப்படுவதில்லை. இங்கு கோபம், வீரம், தாபம், எழுச்சி போன்ற உணர்ச்சிகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழப்பெயர்ப்பு பணிகளுக்கு பொதுவாக அறிவியல் மற்றும் தொழில் நுட்பத்தை சார்ந்த படைப்புகளே எடுத்துக் கொள்ளப் படுகின்றன. மேற்கூறிய இரண்டு கணிப்பொறி வழி மொழி பெயர்ப்பிற்கு எல்லைகளாக கொள்ளப்படுகின்றன.

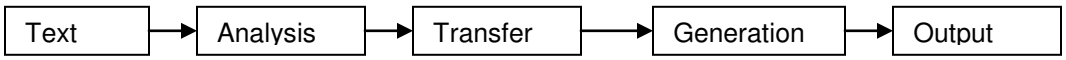
2.18.4. உருஷ்யன்-தமிழ் மொழி பெயர்ப்பு ஒரு கண்ணோட்டம்

இரு மொழிகளுக்கு இடையே உள்ள சொல்வளம் (vocabulary), இலக்கணம் (grammar), கருத்தாடல் (context Structure) இவற்றின் ஒற்றுமை வேற்றுமைகளை ஆராய்வது வகைப்பாடடியல் ஆகும். இரு மொழிகளுக்கு இடையே உள்ள வேற்றுமைகளை மட்டும் அறிந்து ஆராய்வது வேற்றுமை பாராட்டியாகும். உருஷ்யன்-தமிழ் மொழிபெயர்ப்பு பணியில் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமைகள் ஆராயப்பட்டு தொடங்கப் பெற்றது. இவ்விரு மொழிகளுக்கு இடையே உள்ள சொல்வளம், சொற்களுக்கு இடையே உள்ள உறவுகள், தொடரியல் மற்றும் வாக்கியங்களுக்கு இடையே உள்ள உறவுகள் கண்டறியப்பட்டு வரைமுறைபடுத்தப்படுகின்றன. இம்மொழி பெயர்ப்பிற்கு பின்வரும் இரு குறிகோள்கள் அடிப்படையாகக் கொள்ளப்பட்டன:

1. எழுத்து அல்லது இலக்கிய உரைகளை மட்டும் மொழி பெயர்ப்பிற்கு எடுத்துக் கொள்வது. பேச்சு மொழி வழக்கினை ஒதுக்குவது
2. இலக்கிய படைப்புகளை தவிர்த்து அறிவியல் மற்றும் தொழில் நுட்ப உரைகளை மட்டும் எடுத்துக் கொள்வது

2.18.4.1. முன்னேற்றப் பணிகள்

இம்மொழி பெயர்ப்புக்கு உருஷ்ய மொழி மூலமொழியாகவும் (Source Language) தமிழ் இலக்கு மொழியாகவும் (Target Language) எடுத்துக் கொள்ளப்பட்டன. மூலமொழி அதாவது ருஷ்யன் மொழியில் அமைந்துள்ள செய்திகளை ஒலிபெயர்ப்பு (translitration) செய்து கணிப்பொறிக்கு உள்ளீடு செய்வது.இம்மொழி பெயர்ப்புக்கு உருஷ்ய மொழியை மூலமொழியாகவும் (Source Language) தமிழை இலக்கு மொழியாகவும் (Target Language) எடுத்துக் கொள்ளப்பட்டன.உருசிய மொழிக்கும், தமிழுக்கும் ஒரு அகராதி உருவாக்கப்பட்டது.உருஷிய மொழியில் அமைந்துள்ள சொற்கள் சொறறொடர்கள் மற்றும் வாக்கிய அமைப்புகளில் சில முன்திருத்தம் (Preprocessing) செய்தல். மூல மொழியில் உள்ள சொல் மற்றும் வாக்கிய அமைப்பிற்கு நிகரான தமிழ்ச்சொல் வாக்கியம் ஆகியவற்றை நிர்ணயம் செய்ய விதிமுறைகளை உருவாக்குவது இத்தகைய விதிமுறைகளும், வழிமுறைகளும் மாற்றம் (Transfer) என்ற பெயரில் பதிவு செய்தல். கணிப்பொறி வழி மொழி பெயர்ப்பு செய்த பின்னர் தேவைப்படின் தமிழ் மொழிக்கேற்ப வாக்கிய அமைப்புகளை திருத்தி அமைப்பது (Post Processing).



உள்ளீடு பகுப்பாய்வு மொழிமாற்றம் உருவாக்கம் வெளியீடு

2.18.4.2 உருசியன்-தமிழ் மொழி பெயர்ப்பின் படிநிலைகள்

உருஷ்யன்-தமிழ் மொழி பெயர்ப்பிற்கான படிநிலைகளின் விளக்கம் கீழே தரப்பட்டுள்ளன.

மூலமொழி உருசியனை L1 என்றும், தமிழ் மொழியை L2 என்றும் குறிப்பிட்படுகின்றன. முதற்கட்டமாக மூலமொழியில் உள்ள செய்தி ஒலிபெயர்ப்பு செய்யப்பட்டு கணிப்பொறிக்கு உள்ளீடு செய்யப்படுகின்றன. 2-ஆம் கட்ட நிலையில் மூலமொழியில் உள்ள செய்தி பகுப்பாய்வு செய்யப்படுகிறது. இந்நிலையில் சொற்களின் இலக்கணக் கூறுகள் கண்டறியப் படுகின்றன. இதற்கு உருஷ்யன்தமிழ் அகராதி உதவுகிறது. இப்பகுப்பாய்வு மூலம் ருஷ்ய மொழியில் உள்ள செய்திகள் NP, UP, N.N, ADI எனப் பகுப்படுத்தப்படுகின்றன.

மூன்றாம் கட்ட நிலையில் பகுப்பாய்வு செய்யப்பட்டுள்ள மூலமொழியின் வார்த்தையின் சொற்களின் வடிவம், வாக்கிய அமைப்பு போன்றவைகளுக்கு இலக்கு மொழியில் நிகராக உள்ள சொற்களின் வடிவம் வாக்கிய அமைப்புன் மூலம் கண்டறியப்பட்டுள்ளது. மொழி மாற்றம் செய்யப் படுபகின்றன. இந்நிலையில் இவ்விரு மொழிகளுக்கும் இடையே உள்ள முரண் அறிவு (Contrast Knowledge) பயன்படுத்தப்படுகின்றன. கருத்தாடலை பொறுத்த வரையில் இவ்விரு மொழிகளுக்கு இடையே ஒன்றுக்கு-ஒன்று என்ற (one–to–one) உறவு உள்ளது. வேற்றுமைகள் அதிகரிக்க அதிகரிக்க சிக்கல்கள் ஏற்படுகின்றன. பெயர் வரிச் சொல்லைப் பொறுத்த வரையில் இவ்விரு மொழிகளுக்கு இடையே பலவற்றிற்கு-ஒன்று என்ற (many–to–one) உறவு உள்ளது. இவ்வுறவினை கீழ்க்கண்ட படம் விளக்குகிறது.

உருசியன் தமிழ்

. ADJ

SN

PL ADJ
CASES
GENDERS

மேலும் இவ்விரு மொழிகளுக்கு இடையே உள்ள வேற்றுமை உறுப்புகளில் ஒன்றிற்கு பல என்ற (one-to-many) உறவு உள்ளது.

உருசியன்	தமிழ்
	KIRAAL – கின்றாள்
ON	KIRAAN – கின்றான்
	KIRAAR – கின்றார்

இது போன்று இவ்விரு மொழிகளுக்கு இடையே உள்ள வேறுபாடுகள் விதிமுறைகளாகவும் வழிமுறையாகவும் மாற்றம் (transfer) என்ற அமைப்பின் கீழ் தரப்பட்டுள்ளது. பகுப்பாய்வு செய்யப்பட்ட மூலமொழி செய்தி அமைப்பின் உள்ளீடாக செல்கிறது. இந்நிலையில் மூலமொழி அமைப்பிற்கு நிகரான இலக்கு மொழி அமைப்பிற்கு செய்தி மொழி மாற்றம் செய்யப்படுகிறது.

நான்காவது கட்ட நிலையில் மொழி மாற்றம் செய்யப்பட்டுள்ள இலக்குமொழி கட்டமைப்புகள் இலக்கு மொழியின் தன்மைக்கேற்ப அமைவதில்லை. இவ்வாறு கிடைக்கப்பெற்ற இடைநிலை மொழி மாற்ற அமைப்புகளை இலக்குமொழியின் தன்மைக்கேற்ப உருவாக்கி தருவது தான் GENARATION என்ற பகுதியின் பணி ஆகும். இந்நிலையில் மொழிமாற்றம் செய்யப்பட்ட செய்திகள் உள்ளன. இச்செய்தியை மூலமொழியின் வரிவடித்திலோ அல்லது ஒலிபெயர்ப்பு வடிவத்திலோ தரும் பணியை output என்ற அமைப்பு தருகிறது. மொழிபெயர்ப்பு செய்யப்பட்ட செய்தியைக் காட்சித் திரையில் காணலாம். உருஷியன்-தமிழ்மொழி பெயர்ப்பைப் பொறுத்த வரையில் வெளியீட்டை தமிழ் எழுத்துகளில் காண்பதற்கும் அச்சிட்டுக் கொள்வதற்கும் வசதிகள் உள்ளன. அதுபோல் ஒலிபெயர்ப்பு முறையில் விடையை பெறலாம்.

2.18.5. கணிப்பொறியின் மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழி பெயர்ப்பு பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற போது கணிப்பொறி சிறப்புத் தன்மைகளான அதிவேகம், மிகுந்த நினைவகக் கொள்திறன், தகவல்தளம்

உருவாக்கும் வசதிகள் போன்றவைத் திறமையாகப் பயன்படுத்தப்படுகின்றன. ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி பெயர்ப்பு செய்யப்படும்பொழுது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வடிவமைக்கப்படுகின்றன. எனவே இருமொழி மொழிபெயர்ப்பு பணிக்கு கூட ஒருவருக்கு இருமொழிகளிலும் திறமை இருக்க வேண்டும் என்ற அவசியம் இல்லை. இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணிகளில் பகுப்பாய்வு சிக்கல்கள் உள்ளன. ஒரு மொழியின் வளம், தன்மை ஆகியவற்றை கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையானப் பணி அல்ல. எனவே கணிப்பொறி வழி மொழிபெயர்ப்பு திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப் பெற்றிருப்பினும் இன்னும் முழுமைப் பெறவில்லை. மேலும் கணிப்பொறி அறிவியலின் தொழில் நுட்ப வளர்ச்சி கணிப்பொறியில் மொழி பெயர்ப்பு பணிக்கு மிகவும் பயனுள்ளதாக அமைகிறது. இத்தகைய பணிகளின் குறை நிறைகளை கீழே காண்போம்.

மொழிகளுக்கு இடையே உள்ள சொல் வளம் (வடிவம்), இலக்கணம், கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிகளுக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிப்பெயர்ப்பு பணியின் சிக்கல்கள் குறையும்.

2.18.5.1. உருபனியல் சிக்கல்

உருபனியல் அமைப்பை பொறுத்த வரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

1. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து
2. உருபன்களின் கட்டமைப்பொறுத்து

ஒரு சொல்லில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் மட்டும் இருக்குமேயானால் அவைகளை எவ்வளவு துல்லயமாகப் பிரிக்க இயலும் என்பதை பொறுத்து ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தால் அந்த மொழியை தனிநிலை உருபன் கட்டு (isolating) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டு (polysynthetic) மொழி என்கிறோம். சிபேரியன் மொழி பன்நிலை

உருபன் மொழிக்கு சிறந்த உதாரணம் ஆகும். வியட்நாமில் மற்றும் கேண்டுனஸ் ஆகிய மொழி தனிநிலை உருபன் கட்டு மொழிக்கு சிறந்த உதாரணம் ஆகும். எனவே இத்தகைய மொழிகளுக்குள் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்றபொழுது மொழிகளின் உருபனியல் பற்றிய விரிவான தகவல் அவசியமாகும்.

2.18.5.2. தொடரமைப்புச் சிக்கல்

மொழிகளின் தொடரியல் அமைப்பை (Syntactic Structure) பொறுத்து மொழிகள் என மூன்று வகைகளாகப் பிரிக்கப்படுகின்றன:

1. SVO அமைப்பு

ஜெர்மன், பிரெஞ்ச், ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் இந்த அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரிஸ், அரேபி, ஹிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருபுகள் (preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிப் பெயர்ப்புத் திட்டங்களுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரியல் *வீடு* என்பது தலைச்சொல் (head word) ஆகும். *இராமன்* என்பது சார்பு சொல் ஆகும். தலைச் சொல்லுக்கும், சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை *இன்* என்ற ஒட்டு (affix) விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச்சொல்லுக்கும் இடையே உள்ள உறவை வெளிப்படுத்தும் முறையைப் பொறுத்து மொழியைத் தலைக் குறித்தல் மொழி (head marking languge) சார்புக் குறித்தல் மொழி (dependent marking language) என இரு வகையாகப் பிரிக்கப்படுகிறது. தலைக்குறித்தல் மொழியின் ஒட்டு சார்புச் சொல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி

சார்புக் குறித்தல் மொழியை சார்ந்ததாகும். ஹங்கேரிய மொழி தலைக்குறித்தல் மொழியை சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு இதை விளக்கும்:

ENGLISH	: Man's house
Hungarian	: The man house his
H	: Head word (தலைச்சொல்)
The Man	: Dependent Word (சார்புச்சொல்)
A	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணியை மேற்கொள்கின்றபோது தலைச்சொல்லுக்கும் சார்புச்சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப்பெறுகின்றன என்பது பற்றியெல்லாம் ஆராய வேண்டும்.

2.18.5.3. சொல்வகைப்பாடு பொருத்தச் சிக்கல்

மூலமொழியிலுள்ள ஒரு சொல்லின் வகைபாடும் (part-of-speech) அச்சொல்லுக்கு நிகராக இலக்கு மொழியலுள்ள சொல்லின் வகைபாடும் ஒன்றாக இருக்க வேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக, ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழிப்பெயர்ப்பு செய்யப்படும்போது மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்க வேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்பு பணியில் உள்ள சிக்கல்கள் ஆகும்.

2.18.5.4. கருத்தாடல் சிக்கல்

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கான தளம் விளக்கமானதாகவோ சுருக்கமானதாகவோ இருக்கலாம். எடுத்துக்காட்டாக, சகோதரன் என்ற உறவைக் குறிக்க ஆங்கிலத்தில் brother என்ற சொல் உள்ளது. மூத்த சசோதரனை elder brother என்றும் சொல்கிறோம். இவ்வாறு வயதில் மூத்தவர், இளையவர் என்பதைக் குறிக்க elder, younger என்ற சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் அண்ணன் என்றும், தம்பி என்றும் தனித்தனிச் சொற்கள் உள்ளன. இங்கு சகோதரன் என்ற உறவு மேலும் வகைப்படுத்தப்பட்டு முதியவர், இளையவர் என்ற வகையில் அண்ணன், தம்பி என்ற இரு சொற்கள் உள்ளன.

ஆனால் அது ஆங்கிலத்தில் இவ்வாறான வகை என்பது இல்லை. உறவை வேறுப்படுத்த elder, younger என்ற சொற்கள் பயன்படுத்த நேரிடுகிறது.

2.18.5.5. கலாச்சார வேறுபாட்டுச் சிக்கல்

மூலமொழியிலுள்ள செய்திகளின் கலாச்சாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்குமொழிக்கும் பொருத்தமானதாக அமைதல் வேண்டும் அல்லது இலக்கு கலாச்சார பண்புகளுக்கு ஏற்ப மாற்றித் தக்க அளவு இருத்தல் வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சாரத் தொடர்புகள் மொழிபெயர்ப்புச் செய்ய இயலுமா?, இயலாத? என்பதைத் தெளிவு படுத்தும் . மேலும் சபீர்-வீ.ப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ அல்லது குறைப்பதாகவோ உள்ளது. எனவே மொழிப்பெயர்ப்பில் இத்தகைய உளவியல் ரீதியான சிக்கல்கள் உள்ளன.

2.18.5.6. தொடரியல் பகுப்பாய்வுச் சிக்கல்

தொடரியல் பகுப்பாய்வு (parser) பலப் பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிப்பெயர்ப்பிற்காக பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்கு தொடரின் அல்லது வாக்கியத்தின் பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வுச் செய்யப்படுகிறது. எனவே இயந்திர மொழிப்பெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது. பொருளை சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அதிக அடிப்படை (Knowledge Base) தேவைப்படுகிறது.

2.18.5.7. அகராதிச் சிக்கல்

மொழிப்பெயர்ப்பு பணிகளுக்காக உருவாக்கப் படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிற வகை அகராதிகளை விட துறைச் சார்ந்த அகராதிகள் மொழிப்பெயர்ப்பு பணியை எளிமைப் படுத்துகிறது. மூல மொழியிலுள்ள செய்தியை பொறுத்து அதன் தொடர்புடைய அகராதியைப்

பயன்படுத்திக் கொள்ளலாம். மொழிப்பெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத்தொடர்கள், மேலும் அடிக்கடி பயன்படுத்தக் கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்பட வேண்டும்.

2.18.5.8. பிற சிக்கல்கள்

இயந்திர வழி மொழிபெயர்ப்பு செய்கின்றபொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்கு கூடுதல் தகவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கல்களைப் புரிந்துக் கொண்டு அதற்கேற்ற தீர்வு உரிய முறையில் வழங்கப்பட வேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்க வேண்டியது அவசியமாகும். பெரும்பாலும் படைப்பாற்றல் திறன் செய்திகளில் வெளிப்படுத்தும்பொழுது கணிப்பொறிக்கு கூடுதல் செய்திகள் தேவைப்படுகின்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிப்பெயர்ப்புகளில் உள்ளீடு செய்திகள் முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும், இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. இதை இருவகைகளில் செய்யலாம்:

அ) மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிப்பெயர்ப்பது. பின்னர் மொழிப்பெயர்ப்பை படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகிறதா? என்பதை ஆராய்வது, அப்படி இல்லை மூலமொழிச் செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிப்பெயர்ப்பு செய்வது சரியான கருத்து வெளிவரும் வரை இத்தகைய ஆய்வை மீண்டும் செய்வது.

ஆ) மொழிப்பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து தேவையானத் திருத்தங்களை செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தையும் மிச்சப்படுத்துகிறது. மொழிப்பெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிப்பெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப் பணி திறமையானதாக அமைகிறது.

கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிஎன்பது ஒரே மூச்சில் ஒட்டு மொத்தமாகச் செய்யக் கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்பட வேண்டும். அதன் செயல் திறனை அறிய வேண்டும். பணியின் தன்மையை மதிப்பிட வேண்டும். இதற்கு ஏற்றாற் போல் தேவைப்படின் மென்பொருள் கூடுதல் வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை மனிதனுக்கு நிறைவழிக்கும் செயல்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்பு கடுமையானப் பணியாகும். இதற்கு மனித மொழிப்பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (models) வழி முறைகளும் அவசியமாகும். கணிப்பொறி மொழிப்பெயர்ப்பு எண்ணற்ற முயற்சிகளினால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமானப் பணியாகும். இங்கு அறிவியல் மட்டுமல்லாது மொழிப்பெயர்ப்பிற்கு தேவையான கலைத் திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் உருவாக்கப்பட்டுள்ள மொழிப்பெயர்ப்பு பணிகள் மனதிற்கு நிறைவழிக்கக் கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிப்பெயர்ப்பிற்கு உதவுகின்றன. இவைகளைச் செம்மைப்படுத்தி முற்றிலும் சிறந்ததாக அமைப்பதற்கு முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. எனவே தற்பொழுது கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிகள் வரைவு மொழிப்பெயர்ப்பிற்கு போதுமானதாகும்.

மனிதனின் படைப்பாற்றல், கற்பனைத்திறன், உளவியல் பாங்கு போன்றன செய்திகளில் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணியில் இத்தகைய கூறுகளை சேர்க்க வேண்டி உள்ளது. இதற்கு மனித நுண்ணறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன. இத்திட்டங்களில் வெற்றி பெறுகின்றபொழுது திறமையான எதிர்ப்புகளுக்கு ஏற்ற மொழிப்பெயர்ப்பு மென்பொருள்கள் உருவாக்கப்படும். அவ்வகையில் மொழிப்பெயர்ப்பிற்கு உதவக் கூடிய வரை நிலையான மொழிப்பெயர்ப்பையே கணிப்பொறி வழி கிடைக்கப்பெற இயலும்.

2.18.6. தமிழில் இயந்திர மொழிபெயர்புக்கான ஆயத்தங்கள்

பகுதி 3

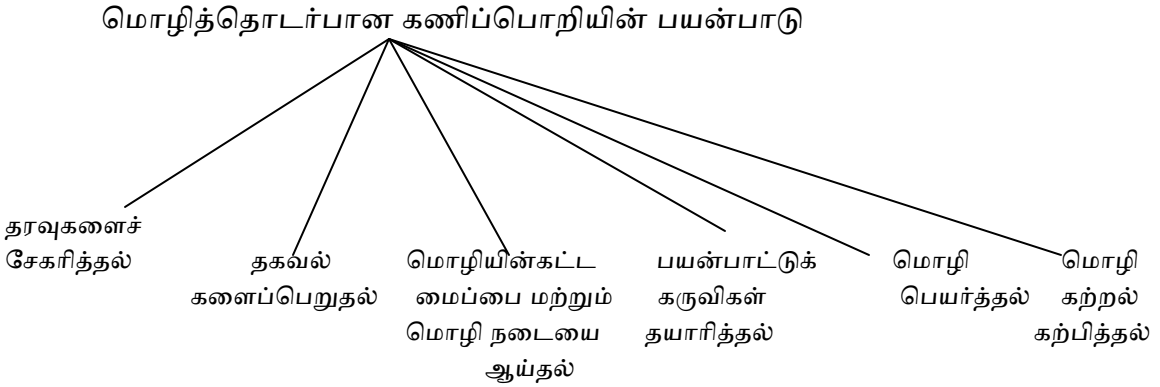
முடிவுரை

மேற்கண்ட பகுதிகள் மொழி ஆய்வில் கணிப்பொறியின் முக்கியத்துவத்தையும் தமிழில் இயற்கை மொழி ஆய்வின் பரிமாணத்தையும் நமக்கு வெளிப்படுத்துகின்றன. கணிப்பொறி மொழியியல் ஆய்வுகளுக்கு ஒரு வரப்பிரசாதமாய் மாறியுள்ளது. மொழியியல் ஆய்வுகள் கணிப்பொறி அறிவியலுக்குத் தேவை என்ற நிலையுடன் கணிப்பொறி மொழியியல் ஆய்வுகளுக்கு மிகத் தேவை என்ற நிலையும் இன்றைய கால கட்டத்தில் நிலவுகின்றது. கணிப்பொறி மொழியின் அமைப்பை ஒலியனியல், உருபனியல், தொடரியல், பொருண்மையியல் என்ற மட்டங்களில் ஆய்வதற்கு உதவுவதுடன் மொழிக்கருவிகள் உருவாக்குவதற்கும் இன்றியமையாத சாதனமாய் பயன்படுகின்றது. அட்டைகளில் பதிவுகளையும் பொருள்களையும் எழுதும் காலம் போய் அகராதி தயாரிக்கும் மென்பொருள்களைப் பயன்படத்தி அகராதிகள் தயாரிக்கும் காலம் நிலவுகின்றது. கணிப்பொறியின் வரவால் அகராதி தயாரிப்பது ஒரு தண்டனையாக இல்லாமல் இனிய அனுபவமாக மாறிவிட்டது. ஆசிரியரின் இடையீட்டால் உற்சாகமில்லாத செயல்பாடாக இருந்த மொழி கற்றல் கணிப்பொறியின் உதவியால் நினைத்தால் இனிக்கும் அனுபவமாக மாறியுள்ளது. கற்றல் மென்பொருள்களைப் பயன்படுத்தி குழந்தைகள் உற்சாகமாக விளையாடிக்கொண்டே மொழியைக் கற்கின்றனர். அகரவரிசை அட்டவணை படுத்துவதற்கும் சொல் அடைவுகள் தயாரிப்பதற்கும் அட்டைகளைப் பயன்படுத்துவதைக் கணிப்பொறி இல்லாமல் செய்துவிட்டது. உரையிலுள்ளதை பேச்சாகவும் பேச்சை உரையாகவும் மாற்றும் மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. எழுத்துணரியின் வரவால் கணிப்பொறியில் தட்டச்சுசெய்து உரைகளை உள்ளீடு செய்யவேண்டும் என்ற நிலை மாறியுள்ளது. கடுமையான வேலையாக இருந்த மொழிபெயர்ப்பு கணிப்பொறியின் பயன்பாட்டால் எளிய வேலையாக மாறியுள்ளது. எழுத்துரையை மட்டுமன்றி பேச்சுரையையும் உடன்தானே மொழிபெயர்க்கும் VERB MOBILE போன்ற கருவிகள் உருவாக்கபட்டு பயன்பாட்டிலுள்ளன. இனி வரும் காலத்தில் செல்போனில் ஒருவர் ஆங்கிலத்தில் பேசினால் மற்றவர் அதைத் தமிழில் கேட்கவியலும். இதுவரை முன்னேற்றத்தின் பரிமாணத்தையும் வேகத்தையும் கணிக்க முடிந்தது. கணிப்பொறியின் பயன்பாட்டால் முன்னேற்றத்தின் பரிமாணத்தையோ வேகத்தையோ கணிக்கவியலாத சூழல் உருவாகியுள்ளது. நாளைய உலகம் கடவுள் கையில் என்று கூறவியலாமல் கணிப்பொறி கையில் என்று கூறத்தோன்றுகின்றது.

மொழி ஆய்வில் கணிப்பொறியின் பயன்பாடு
(Use of Computer in Linguistic Research)

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப்பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களைப் பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. கணிக்கும் பொருளாக மட்டும் இருந்த கணிப்பொறி, தரவுகள் சேமித்து வைக்கும் கருவியாக, தகவல் தொடர்புக் கருவியாக, அத்தகவல்களைப் பயன்படுத்திக் கொள்ள உதவும் கருவியாக வளர்ந்துள்ளதோடு, நாம் தரும் தரவுகளைத் தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகின்றது. மொழிச்செயல்பாட்டிலும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

மொழித் தொடர்பான கணிப்பொறியின் பயன்பாட்டை ஆறு முக்கியக் கூறுகளாகப் பகுக்கலாம்.



* இணைப்பேராசிரியர் மற்றும் துறைத்தலைவர், மொழியியல் துறை, தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்-613 005.

முதல் இரண்டும் மொழி அல்லாத பிற நெறிகளுக்கும் பொருந்தும். மூன்றாவது பயன்பாடு பின்னர்வரும் மூன்று பயன்பாட்டிற்கும் அடிப்படையாக அமையும்.

தரவுகளை சேகரித்தல்

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவிலுள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கான சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில்

(நினைவகம், பதிவுத் தட்டு, பதிவு நாடா போன்றவற்றில்) சேகரிப்பு வைத்துத் தேவையான போது பயன்படுத்திக்கொள்ள இயலும்.

மொழிக்கான தரவுகள் பல வகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவொ வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். ஒரு பனுவலாக (Text) அமையலாம். (எந்த ஒரு மொழி உடலும் பனுவலாகும்.) வெவ்வேறு வகைப்பட்ட பனுவல்களில் இருந்து தெரிந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு பெருந்தரவு (Corpus) என்று அழைக்கப்பெறும். தமிழுக்கான பெருந்தரவைச் சேகரித்து வைப்பது மிக இன்றியமையாத பணியாகும். பல இலக்கியக் கூறுகளை உள்ளடக்கிய பனுவல்கள், பல நடையியல் கூறுகளை உள்ளடக்கிய பனுவல்கள், வட்டார வழக்கு, கிளைமொழி வழக்கு, சமுதாய மொழி வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய பனுவல்கள் எல்லாம் ஒரு சேரப் பெருந்தரவாக அமையும். இத்தகைய பெருந்தரவை இருவகையாகப் பிரிக்கலாம். ஒன்று எழுத்து வடிவில் அமையும் பெருந்தரவு (Written Corpus) மற்றொன்று பேச்சு வடிவில் அமையும் பெருந்தரவு (Speech Corpus). இத்தகைய பெருந்தரவுகள் மொழி ஆய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவி புரியும்.

தகவல்களைப் பெறுதல்

தரவுகள் கணிப்பொறியால் ஆயப்பட்டுத் தகவல்களாகத் தரப்படும். மொழித்தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வன் தட்டு, குறுந்தட்டு, பதிவு நாடா போன்றவற்றில்) சேமித்து வைத்துத் தேவைப்படும் போது பயன்படுத்திக்கொள்ளலாம். பல்வேறு விதமான தகவல்கள் இணையதளத்தில் பெருமளவில் நிரம்பிக் கிடக்கின்றன. இத்தகவல்களைப் பெறுவதில் முக்கியமான அம்சம் என்னவென்றால் இயல் மொழியைப் பயன்படுத்தித் தேவையான தகவல்களைப் பெறுதலாகும். இதற்குத் தேவை இயல் மொழி இடைமுகப்பு (Natural Language Interface). இதனால் இயல் மொழி அடிப்படையிலான குறியீடுகளைக் கொண்டே தேவையான தகவல்களைப் பெற இயலும். தமிழ்

மொழியின் கட்டமைப்பைக் கணிப்பொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ்மொழியைப் பயன்படுத்தித் தகவல்களைப் பெற உதவும் இயல்மொழி இடைமுகப்புகளை நிறுவ இயலும். இத்தகைய இடைமுகப்பு பல்வேறு வடிவுகளைக் கொண்டிருக்கும். தகவல் தேடும் பொறி (Search Engine) தகவல்களைப் பெற உதவும் முக்கிய கருவியாகும். தமிழுக்கு இத்தகைய தகவல் தேடும் பொறி உருவாக்குவதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டும் அவை பெருமளவில் வெற்றியும் அடைந்துள்ளன.

மொழியின் கட்டமைப்பை ஆய்தல்

மொழிப்பனுவல்களைப் படிநிலை (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழியலகுகளைப் பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்துகொள்ளக் கணிப்பொறியைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக்கொள்ளலாம்.

பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்.

பள்ளிமாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான் என்பது வினைத் தொடர் என்றும் மேலும் பள்ளி, மாணவன், மரம், பலாப்பழம், ரகுமான் என்பன பெயர்கள் என்றும், பறி என்பது வினை என்றும் பெரிய என்பது பெயரடை என்றும் கவனமாக என்பது வினையடை என்றும் இலிருந்து, ஐ என்பன வேற்றுமை உருபுகள் என்றும், இலிருந்து, த என்பது இறந்த கால இடைநிலை என்றும், ஆன் என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிப்பொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுத்தாய்வு செய்து தந்து விடும்.

மொழிப் பனுவலை இடுவராலாகத் (Input) தந்து தக்க வழியமைப்புகளும் (Programs) கட்டளைகளும் தந்தால், மொழியின் கட்டமைப்பு குறித்த தகவல்களை விடுவரலாகப் (Output) பெறலாம். கணிப்பொறியைக் கொண்டு

மொழிப்பனுவல்களை வாக்கியங்களாகவும் வாக்கியங்காளைத் தொடர்களாகவும் தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக்கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளுக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்காகத் தொடரியல் ஆய்வி (Syntactic Analyzer), உருபனியல் ஆய்வி (Morphological Analyzer), ஒலியனியல் ஆய்வி (Phonological Analyzer), பொருண்மையியல் ஆய்வி (Semantic Analyzer) என்ற ஆய்விகளைத் தயாரிக்க வேண்டும். தமிழில் இத்தகைய ஆய்விகள் தயாரிக்கப்பட்டுப் பயன்பாட்டில் உள்ளன.

மொழி நடையை ஆய்தல்

மொழியின் கட்டமைப்பைக் கருத்தாடல், வாக்கியம், தொடர் சொல் என்ற அலகுகளின் அடிப்படையில் கணிப்பொறி மூலம் ஆய இயலும் என்று கண்டோம். மொழி அலகுகள் நடையியல் அடிப்படையிலான சிறப்புப் பண்புக் கூறுகளை வெளிப்படுத்தி மொழிநடை வேறுபாட்டை உணர்த்தும். கணிப்பொறியின் உதவியால் வேறுபட்ட நடையில் அமைந்த பனுவல்களை அல்லது மொழி வழக்குகளை ஆய்ந்து ஒப்பீடு செய்து அவற்றின் நடையியல் சிறப்புப் பண்புகளைத் தெரிந்து கொள்ளலாம். இத்தகைய ஆய்வு படைப்பாளியை இனங்கண்டு கொள்ளவும் படைப்பிலக்கியத்தின் காலத்தைக் கணிக்கவும் பயன்படும். கணிப்பொறி உதவியுடன் இத்தகைய ஆய்வை நுணுக்கமாகவும் விரைவாகவும் செய்து முடிக்கலாம். இத்தகைய, கணினி வழி தமிழ் மொழி நடை ஆய்வுகள் பல வெற்றிகரமாகச் செய்து முடிக்கப்பட்டுள்ளன. புள்ளியல் மொழியியல் (Statistical Linguistics), அளவீட்டு மொழியியல் (Quantitative Linguistics) போன்ற பயன்பாட்டு மொழியியல் பாடங்களின் கோட்பாடுகளும் கொள்கைகளும் மொழியியல் நடை ஆய்வுக்கு உறுதுணை புரிகின்றன. மொழிநடையை ஆய உதவும் மென்பொருள்கள் பல கண்டு பிடிக்கப்பட்டு மொழிப் பயன்பாட்டில் உள்ளன.

பயன்பாட்டுக் கருவிகள் தயாரித்தல்

சொற்கள் தொடர்பான பயன்பாட்டுக் கருவிகள்

சொற்கள் தொடர்பான பயன்பாட்டுக் கருவிகளைத் தயாரிப்பதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். பின்வருவன இதில் அடங்கும்.

1. சொற்களை அகரவரிசைப்படுத்தல்
2. சொல்லடைவு தயாரித்தல்
3. சொல்லாட்சி அகராதி மற்றும் பொருளடைவு தயாரித்தல்
4. அகராதி தயாரித்தல்
5. சொற்களஞ்சியம் தயாரித்தல்
6. எழுத்து மற்றும் இலக்கணப் பிழை திருத்தி தயாரித்தல்
7. சொல் ஆய்வி தயாரித்தல்

சொற்களை அகரவரிசைப் படுத்தல்

தமிழ்மொழியின் அகரவரிசை தொடர்பான செய்திகளையும் வழியமைப்பு ஆணைகளையும் (Program Instructions) கணிப்பொறிக்குத் தந்தால் கணிப்பொறி மிகக் குறைந்த நேரத்தில் வேண்டிய சொற்களை அகர வரிசைப் படுத்தித் தந்துவிடும். ஆங்கிலத்திற்கு இது மிகச் சாதாரண ஒன்று. தமிழைப் பொறுத்தவரையில் பலவிதமான எழுத்துருக்களும் (Fonts) அதற்குப் பின்னணியாய் அமையும் வேறுபட்ட குறியன்களும் (Codes) குறியன் எண்களும் (Code Numbers) பயன்படுத்தப்படுவதால் அகரவரிசைப் படுத்தல் எளிய செயலாக அமையவில்லை. இருப்பினும் வேறுபட்ட எழுத்துருக்களை வழியமைப்புகள் மூலம் இணைத்து அகரவரிசைப்படுத்தும் மென்பொருள்கள் கண்டுபிடிக்கப்பட்டுப் பயன்பாட்டில் உள்ளன.

சொல்லடைவு தயாரித்தல்

சொல்லடைவு தயாரிப்பதையும் கணிப்பொறி மூலம் விரைவில் செய்துவிட இயலும். சொற்களையும் அவற்றோடு தொடர்பான செய்திகளையும் தந்தால் குறைந்த கால நேரத்தில் கணிப்பொறியின் உதவியால் சொல்லடைவு தயாரித்து விடலாம்.

எடுத்துக்காட்டாகச் சங்க இலக்கியங்களுக்குச் சொல்லடைவு தயாரிப்பதை எடுத்துக் கொள்வோம். பின்வரும் அடிகள் சங்க இலக்கியமான புறநானூற்றிலிருந்து எடுக்கப்பட்டவை.

1. இவனியாரென்குவை யாயி னிவனே (புறம் 83.1)
2. தூங்குகையா னோங்க் நடைய (புறம் 22.1)

இவ்வடிகள் சொற்களாகப் பின்வருமாறு பிரித்துத் தரப்படவேண்டும். வழியமைப்புகள் மூலம் இச்சொற்களை அகரவரிசைப்படுத்தி அவை ஒவ்வொன்றும் வரும் இடங்களைக் குறிப்பிட்டுத்தர இயலும்.

ஆயின் (புறம் 83.1)
இயார் (புறம் 83.1)
இவன் (புறம் 83.1)
என்குவை (புறம் 83.1)
ஏ (புறம் 83.1)
ஓங்கு (புறம் 22.1)
கையால் (புறம் 22.1)
தூங்கு (புறம் 22.1)
நடைய (புறம் 22.1)

ஒரு சொல் சங்க இலக்கியத்தில் எத்தனை முறை கையாளப் பெறுகின்றது அல்லது குறிப்பிட்ட ஒரு நூலில் எத்தனை முறை கையாளப் பெறுகின்றது என்பன போன்ற தகவல்களை வழியமைப்புகள் மூலம் பெறலாம். தமிழ்ப்பல்கலைக்கழக அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்க இலக்கியங்களுக்கான ஒரு சொல்லடைவைத் தயாரித்துள்ளன.

சொல்லாட்சி அகராதி மற்றும் சொற்பொருளடைவு தயாரித்தல்

சொல்லாட்சி அகராதி (Concordance) ஒரு பனுவலில் இடம் பெறும் சொற்களை அகரவரிசைப்படுத்தித் தருவதோடு அச்சொற்கள் நூலில் இடம்பெறும் இடங்கள், எத்தனை முறை இடம்பெறுகின்றன என்ற கணக்கீடு, எந்தெந்தச் சூழலில் பயன்படுத்தப்படுகின்றன என்பன போன்ற செய்திகளைத் தரும். சொற்பொருளடைவு (Compendium) சொல்லாட்சி அகராதிகளினின்றும் ஒருபடி மேற்சென்று சொற்கள் வரும் இடங்களையும் சூழல்களையும் பொருள்

அடிப்படையில் பிரித்துத் தரும். கணிப்பொறியைப் பொறுத்தவரையில் இரண்டையும் ஒன்றில் அடக்கித் தந்துவிடலாம்.

புறநானூற்றிலிருந்து எடுக்கப்பட்ட ஒரு சொல்லின் சொற்பொருளடைவின் மாதிரி கீழே தரப்பட்டுள்ளது.

அணி

1. அழகு

அழகு தாழ்ந்த அணி (புறம் 57.1)
அணிகொள் தேர் (புறம் 68.14)

. . .

. . .

. . .

2. அணிகலன்

இழையணிபொலிந்த . . . . . .
(புற, 81.1)

. . .

. . .

. . .

3. ஒப்பனை

ஆடு கோட்பாட்டுச் சேரலாதன் போந்தைப் பொழியின் கண்
அணியல் (படி 51.9)

4. அணிமை
5. வரிசை
6. திரட்சி
7. ஆடை
8. ஒப்புமை
9. சூழ்தல்

10. பக்கரை (சேணம்)

சொல்லடைவு தயாரிக்க முன்னர்க் கூறியது போல் செய்யுளின் அடிகளைச் சொற்களாகப் பிரித்துத் தரவேண்டும். இத்துடன் சூழல் அடையாளப் படுத்தப்பட வேண்டும். வழியமைப்புகள் மூலம் நாம் சொல்லாட்சி அகராதியையோ சொற்பொருடைவையோ தயாரிக்க இயலும். தமிழ்ப்பல்கலைக்கழக் அகராதியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து கணிப்பொறியைப் பயன்படுத்திச் சங்ககால இலக்கியங்களுக்கு இத்தகைய சொல்லாட்சி அகராதியும் சொற்பொருளடைவும் தயாரித்துள்ளன.

அகராதி தாயாரித்தல்

கணினியைப் பயன்படுத்தி, பலர் உழைப்பால் பல்லாண்டுகாலம் செய்யப்படும் அகராதித் தயாரித்தலை மிகக் குறுகிய காலத்தில் குறைந்த ஆட்களின் சக்தியால் செய்து முடித்துவிடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்தில் மின் அகராதிகளையும் தயாரிக்கலாம். இத்தகைய அகராதி கணிப்பொறி வழி மொழி கற்பித்தல், மொழிபெயர்த்தல் போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயன்படும்.

அகராதி தயாரிப்பதற்கான சில மென்பொருள்கள் வெளிவந்துள்ளன. (எடுத்துக்காட்டாக Tool Box என்ற மென்பொருள் அகராதி தயாரிக்கப் பெரிதும் உதவும்.) அவற்றைப் பயன்படுத்தி அகராதிகள் தயாரிக்கலாம். DATA BASE, DATE BASE MANAGEMENT SYSTEMS போன்றவை அகராதி தயாரித்தலுக்குப் பெரிதும் உதவும். பின்வரும் எடுத்துக்காட்டு ஃபாக்ஸ்புரோ போன்ற வழிமுறைகளைப் பயன்படுத்தி அகராதி தயாரித்தலின் எளிய விளக்கமாகும்.

அகராதியின் ஒவ்வொரு சொற்பதிவிலும் தலைப்புச் சொல், எழுத்துப்பெயர்ப்பு, சொல் வகைப்பாடு, பல்பொருண்மைகள் போன்ற தகவல்கள் அடங்க வேண்டும் என்பதன் அடிப்படையில் தரவு மையத்தை உருவாக்க வேண்டும். கீழ்வரும் எடுத்துக்காட்டில் முக்காற்புள்ளிக்கு இடப் பக்கம் வருவனவற்றைப் பீல்டுகளின் பெயர்களாகவும் முக்காற்புள்ளிக்கு வலப்பக்கம்

வருவனவற்றை அவற்றில் நிரப்பப்பட வேண்டிய செய்திகளாகவும் எடுத்துக்கொள்ளலாம். ஒவ்வொரு சொற்பதிவும் ஒரு பதிவாக அமையும்.

தலைப்புச் சொல்:ஏவு
எழுத்துப்பெயர்ப்பு: eevu

சொல் வகைப்பாடு: செ.பொ.குன்றா.வி.
(செயப்படுபொருள் குன்றா வினை)
பொருள் 1: தூண்டுதல்
எடுத்துக்காட்டு 1: குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான்.
பொருள் 2 : 2. கட்டவிழ்த்துவிடல்
எடுத்துக்காட்டு 2 : அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை அடக்கி விட இயலாது.
பொருள் 3 : அனுப்புதல், செலுத்துதல்
எடுத்துக்காட்டு 3 : ஏவுகணைகளை ஏவிப் பலத்த சேதம் விளைவித்தனர்.

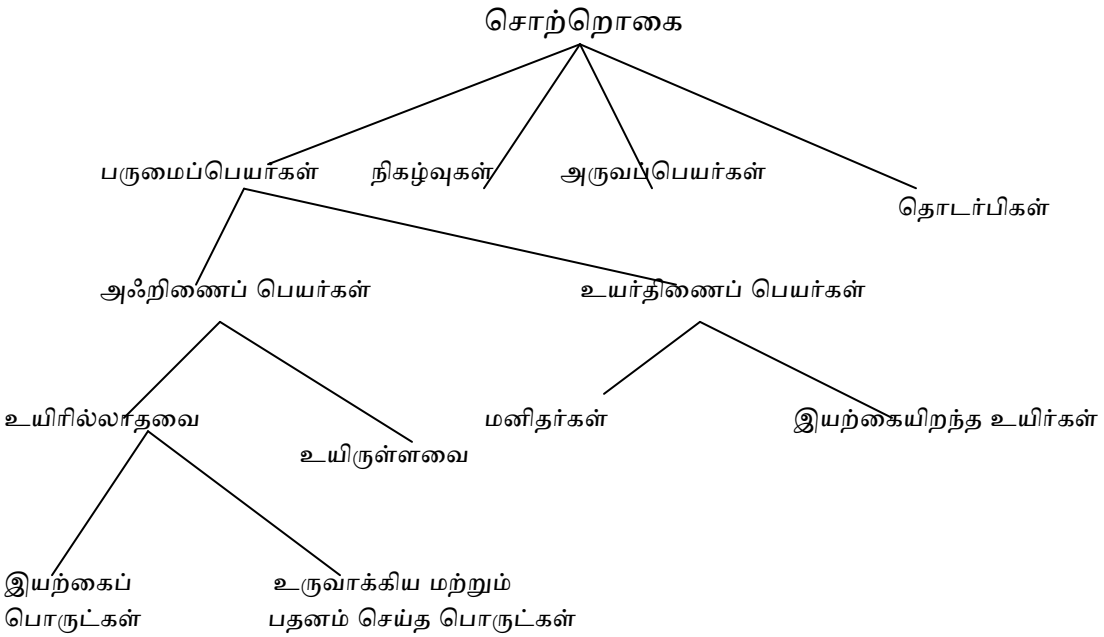
வழியமைப்புகளைப் பயன்படுத்திச் சொற்பதிவைப் பின்வருமாறு அகராதி வடிவில் தரலாம்.

ஏவு. eevu செ. பொ. குன்ற.வி. 1. தூண்டுதல்; குண்டர்களை ஏவி எங்களைப் பயமுறுத்தினான். 2. கட்டவிழ்த்துவிடல்; அடக்கு முறையை ஏவிவிட்டு மக்களின் போராட்டத்தை அடக்கிவிட இயலாது. 3. அனுப்புதல், செலுத்துதல்; ஏவுகணைகளை ஏவிப்பலத்த சேதம் விளைவித்தனர்.

சொற்களஞ்சியம் தயாரித்தல்

சொற்களஞ்சியம் (Thesaurus) தயாரிப்பதற்குக் கணிப்பொறியைப் பயன்படுத்தலாம். ஒத்தல் (Similarity), உள்ளடங்கல் (Inclusion), மேலுறல் (Overlapping), வேறுபடல் (Disjunction) என்ற தர்க்கப் பொருண்மை உறவுகளின் அடிப்படையில் சொற்களை அவற்றின் படிநிலை அமைப்பும் பொருண்மையும் வெளிப்படும் விதத்தில் பொருண்மைக் களங்களாகப் பகுத்துத் தருவது சொற்களஞ்சியமாகும். கணிப்பொறி உதவியுடன் இச்சொற்களஞ்சியத்தைக்

குறைந்த உழைப்பில் குறுகிய காலக் கட்டத்தில் தயாரித்துவிடலாம். கணிப்பொறியில் பயன்படுத்தத்தக்க விதத்தில் மின் சொற்களஞ்சியத்தை உருவாக்கலாம். இத்தகைய மின்சொற்களஞ்சியம் மொழிபெயர்த்தல், மொழி கற்பித்தல் போன்ற செயற்பாடுகளுக்குப் பெரிதும் பயனபடும். தமிழுக்கான மின் சொற்களஞ்சியம் ஒன்றைத் தமிழ்ப் பல்கலைக்கழக மொழியியல் துறையும் கணிப்பொறி அறிவியல் துறையும் இணைந்து தயாரித்துள்ளன.



இப்பாகுபாட்டின் அடிப்படையில் சொற்றொகையைப் பகுத்துப் பொருண்மை உறவுகளின் அடிப்படையில் சொற்களைத் தரவேண்டும்.

1. பருமைப் பெயர்கள்

1.1. அஃறிணைப் பெயர்கள்

1.1.1. உயிரில்லாதவை

1.1.1.1. இயற்கைப் பொருட்கள்

1.1.1.1.1. விண்வெளிப்பொருட்கள்

1.1.1.1.1.1. வானம், ஆகாயம், வான், ஆகாசம், விண், விசும்பு, அந்தரிட்சம், அந்தரம்.

1.1.1.1.1.2. கீழ்வானம்; மேல்வானம்

இது தரவு மையமாக அமையும். இதிலிருந்து வழியமைப்புகள் வழி சொற்களஞ்சியத் தகவல்களைப் பெறலாம்.

சொல்வலை தயாரித்தல்

இது ஒரு இணையதளத் தரவு மையமாகும் (Online Data Base). இது சொற்களஞ்சியம் மற்றும் அகராதி என்ற இரண்டின் பண்புகளையும் கொண்டு மிகுந்த திறன் உடையதாய் அமைந்துள்ளது. சொற்கள் தமக்கிடையே உள்ள பொருண்மை உறவுகளால் வலை போல் பின்னப்பட்டிருப்பதை வெளிப்படுத்துவதாய் இது அமையும். சொல்வலையின் மிகப் பரந்த நோக்கம் சொற்றதகவல்களைச் சொல்வடிவுகள் அடிப்படையில் அல்லாமல் சொற்பொருண்மை அடிப்படையில் ஒழுங்குபடுத்துவதாகும். இந்த அடிப்படையில் சொல்வலை அகராதியிலிருந்து வேறுபட்டுச் சொற்களஞ்சியத்தையும் உள்ளடக்கும். ரொஜஸ்ட் தெசாரஸ் போன்ற சொற்களஞ்சியங்களும் தமிழுக்கு இராசேந்திரனால் (2001) உருவாக்கப்பட்ட தமிழ்ச் சொற்களஞ்சியமும் சொல்வலை பின்னுவதற்கு முன்னோடியாய் அமையும். இவையெல்லாம் புத்தக வடிவில் அமைந்தவை. அகரவரிசையில் அமைந்த சொற்களஞ்சியங்களில் சிக்கல் என்னவென்றால் அவற்றில் தேவைக்கு அதிகமான பதிவுகள் காணப்பெறும். இரு சொற்கள் ஒருபொருள் பன்மொழிகள் என்றால் கூட அகரவரிசைக்காக இரண்டுமுறை நோக்கீடு செய்துதான் தேவையான தகவல்களைப் பெரும்பாலும் பெற இயலும். முதல்முறை அகரவரிசையில் அமைந்த சொல்லடைவு மூலம் குறிப்பிட்ட சொல் எப்பொருண்மைத் தலைப்பில் வருகிறது என்பதைத் தெரிந்துகொண்டு இரண்டாவது முறை நோக்கீடு செய்துதான் வேண்டிய தகவலைப் பெற இயலுகின்றது. இதனால் தேடும் நேரம் இரண்டு மடங்காக உயரும். ஆனால் கணிப்பொறியின் வழி தேடலை விரைவாகச் செய்ய இயலும். சொல் வலை மின்சொற்களஞ்சியத்தைவிட ஒரு படி மேலானது. தமிழ்ப் பல்கலைக்கழகமும் சென்னையிலுள்ள AUKBC Research Centre உம் இணைந்து தமிழ்ச் சொல்வலையைத் தமிழ் இணையப் பல்கலை கழகத்திற்காக உருவாக்கி உள்ளன.

எழுத்து மற்றும் இலக்கணப் பிழை திருத்தி

தமிழைத் தமிழ் எழுத்துருக்களின் உதவியால் கணிப்பொறியில் எழுத இயலும். இக்காலக் கட்டத்தில், கணிப்பொறியில் எழுதப்பட்ட தமிழ்ப்பனுவல்களில் வரும் எழுத்துப் பிழைகளையும் இலக்கணப்பிழைகளையும் கணிப்பொறியே கண்டுபிடித்துத் திருத்தம் செய்வது இன்றியமையாத தேவையாகின்றது. எழுத்து மற்றும் இலக்கணப் பிழை திருத்திட் தயாரிக்கப் பின்வரும் உட்கருவிகள் தேவையாகுகின்றன.

1. கணினித் தொடரியல் ஆய்வி
2. கணினி உருபனியல் ஆய்வி
3. கணினி ஒலியனியல்/எழுத்து ஆய்வி
4. மின் அகராதி

தமிழ்ப்பல்கலைக் கழக மொழியியல் துறையில் இவ்வாய்வு மேற்கொள்ளப்பட்டுள்ளது. தமிழுக்கு எழுத்துப்பிழை திருத்தி ஓரளவுக்குச் செம்மையாக வேலை செய்யும் வகையில் உருவாக்கப்பட்டுள்ளது. இலக்கணப் பிழைதிருத்தி முழுவதுமாக உருவாக்கப்பெறவில்லை. அதற்கான ஆய்வுகள் மேற்கொள்ளப்பட்டு முயற்சிகள் எடுக்கப்பட்டுள்ளன.

தமிழ்ச் சொல்லாய்வி தயாரித்தல்

தமிழுக்கான சொல்லாய்விகள் பல இன்று வெளி வந்துள்ளன. அவை தமிழ்ப்பனுவல்களை உள்ளீடு செய்து வடிவமைத்து வெளியீடாகப் பெறும் நுட்பங்களைக் கொண்டுள்ளன. Word, Word Star போன்ற மென்பொருள்களில் கிடைக்கும் எல்லா வசதிகளும் இவற்றிலும் உள்ளன. சொற்களஞ்சியம், எழுத்து மற்றும் இலக்கணப் பிழைதிருத்தி என்பன மேலும் மேம்படுத்தப்படவேண்டும். பல தமிழ்ச் சொல் சொல்லாய்விகளில் இவ்வசதிகள் இல்லை.

எழுத்துப் பெயர்ப்பு செய்தல்

ஒரு மொழியின் எழுத்துக்களால் எழுதப்பட்ட பிற மொழிகளின் எழுத்துக்களைப் பயன்படுத்திப் படிக்கத்தக்க முறையில் பெயர்ப்பு செய்யும் வசதிமுறைகள் கணிப்பொறியில் வந்துவிட்டன. இதனால் தமிழ் எழுத்தில்

எழுதப்பட்ட பனுவல்களை ரோமன் எழுத்தில் படிக்கவோ ரோமன் எழுத்தில் எழுதப்பட்ட பனுவல்களைத் தமிழ் எழுத்தில் படிக்கவோ இயலும்.

ஒலிபெயர்ப்பு செய்தல்

கணிப்பொறியில் எழுதப்பட்ட பனுவலை ஒலிப்பு வடிவில் வரும்படி செய்தலும் முக்கியமான ஒரு ஆய்வாகும். இத்தகைய ஒலிபெயர்ப்பிகளைப் பனுவலிலிருந்து பேச்சாகப் புரிந்துகொள்ள உதவும் கருவியை Text to Speech Recognizer என்பர். தமிழில் இதற்கான முயற்சிகள் தொடங்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

பேச்சுக்கூட்டிணைப்பான் மற்றும் பேச்சு அறிவான் தயாரித்தல்

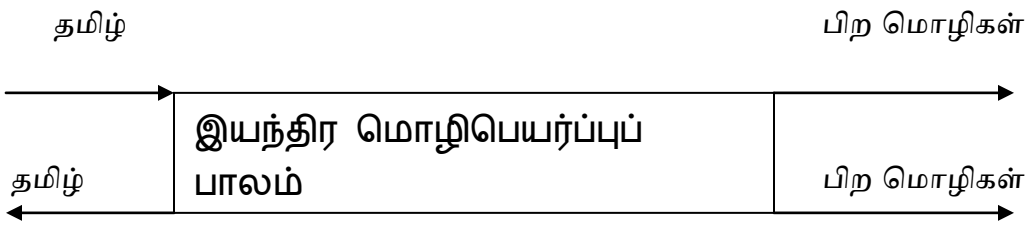
பேச்சுக்கூட்டிணைப்பான் (Speech Synthesizer) மற்றும் பேச்சு அறிவான் (Speech Recognizer) தயார்செய்வது கணிப்பொறியை நேரடிமொழிப் பரிமாற்றக் கருவியாக மாற்றவும் ரோபட் போன்றவற்றை இயல்மொழியின் மூலமும் இயக்கவும் முடியும். இத்தகைய கருவிகள் தயாரிப்பதற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு நடந்து வருகின்றன.

பார்வை எழுத்தறிவான் தயாரித்தல்

அச்சிடப்பட்ட பனுவல்களை அப்படியே கணினிக்கு எழுத்துக்களாகத் தானாகவே உள்ளீடு செய்ய இயலும் கருவியைப் பார்வை எழுத்தறிவான் (Optical Character Recognitior) என்பர். இக்கருவி தட்டச்சு முறைப்படி சாவிப்பலகை வழி பனுவல்களை உள்ளீடு செய்ய ஆகும் கால விரையத்தை இல்லாமல் செய்கின்றது. ஆங்கிலத்திற்கு இத்தகைய கருவி உருவாக்கப்பட்டுப் பயன்படுத்தப்படுகின்றது. ஆனால் தமிழுக்கு அதன் எழுத்துக்களின் கலவைத் தன்மை காரணமாக இத்தகைய கருவியை உருவாக்குவது கடினமான செயலாகும். இருப்பினும் தமிழுக்கு இத்தகைய பார்வை எழுத்தறிவான் உருவாக்கும் முயற்சிகள் எடுக்கப்பட்டு ஓரளவுக்கு வெற்றியும் கிட்டியுள்ளது.

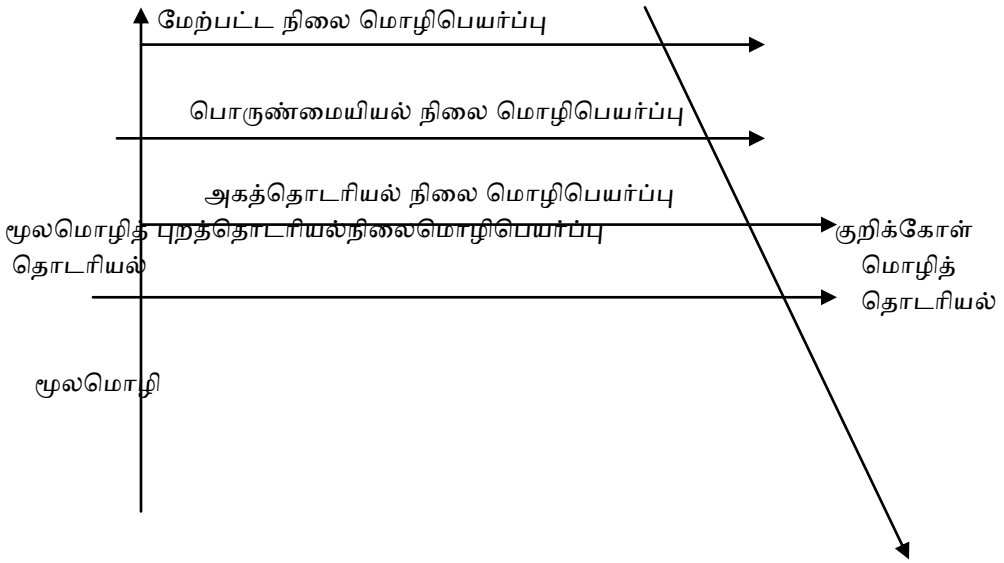
மொழிபெயர்ப்புச் செய்தல்

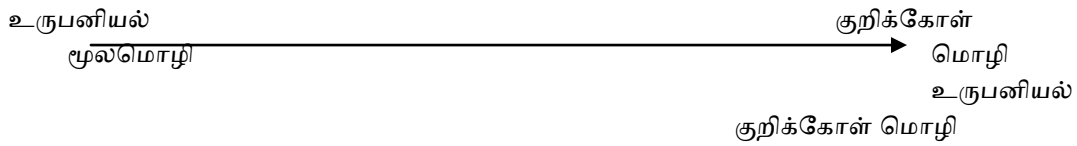
ஒரு மொழியிலுள்ள செய்திகளைப் பிற மொழிக்கு மாற்றித்தருவதை மொழிபெயர்ப்பு என்று சுருக்கமாய்க் கூறலாம். மொழி பெயர்ப்பு செய்வதற்கு மனித மூளைதான் தலைசிறந்த கருவி என்றாலும், மனித மொழிபெயர்ப்பில் ஏற்படும் கால தாமதத்தைத் தவிர்க்க இயந்திர மொழிபெயர்ப்பைக் கைக்கொள்ளலாம். இயந்திர மொழிபெயர்ப்பு செய்வதற்கான இயந்திரம் கணிப்பொறிதான். தமிழிலிருந்து பிற மொழிகளுக்கு மொழிபெய்ர்ப்பதற்கும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு மொழிபெயர்ப்பதற்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் பாலம் அமைத்தால் அது மொழிப் பரிமாற்றத் தடங்கலை உடைத்தெறியும்.



மூல மொழி மற்றும் குறிக்கோள் மொழி இவற்றின் கட்டமைப்புக்குத் தகுந்தவாறு இயந்திர மொழிபெயர்ப்பின் தன்மை மாறும். அதிக ஒற்றுமையுள்ள இரு மொழிகளைச் சொல் அல்லது உருபனியல் நிலையிலேயே நேரடியாக மொழி பெயர்க்கலாம். மொழிகளின் கட்டமைப்பு வேறுபாட்டிற்குத் தகுந்தவாறு, புறத்தொடர் நிலையிலோ அகத்தொடரியல் நிலையிலோ பொருண்மையியல் நிலையிலோ மொழிபெயர்க்கலாம்.

எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்புப் பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.





எந்த அளவிற்கு இயந்திர மொழிபெயர்ப்பில் மனித உழைப்பு பயன்படுகிறது என்பதன் அடிப்படையில் இயந்திர மொழி பெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்.

1. மனித உதவியுடன் நடைபெறும் இயந்திர மொழிபெயர்ப்பு
2. கணிப்பொறி உதவியுடன் நடைபெறும் இயந்திர மொழி பெயர்ப்பு
3. கணிப்பொறி தானே மொழிபெயர்க்கும் இயந்திர மொழி பெயர்ப்பு

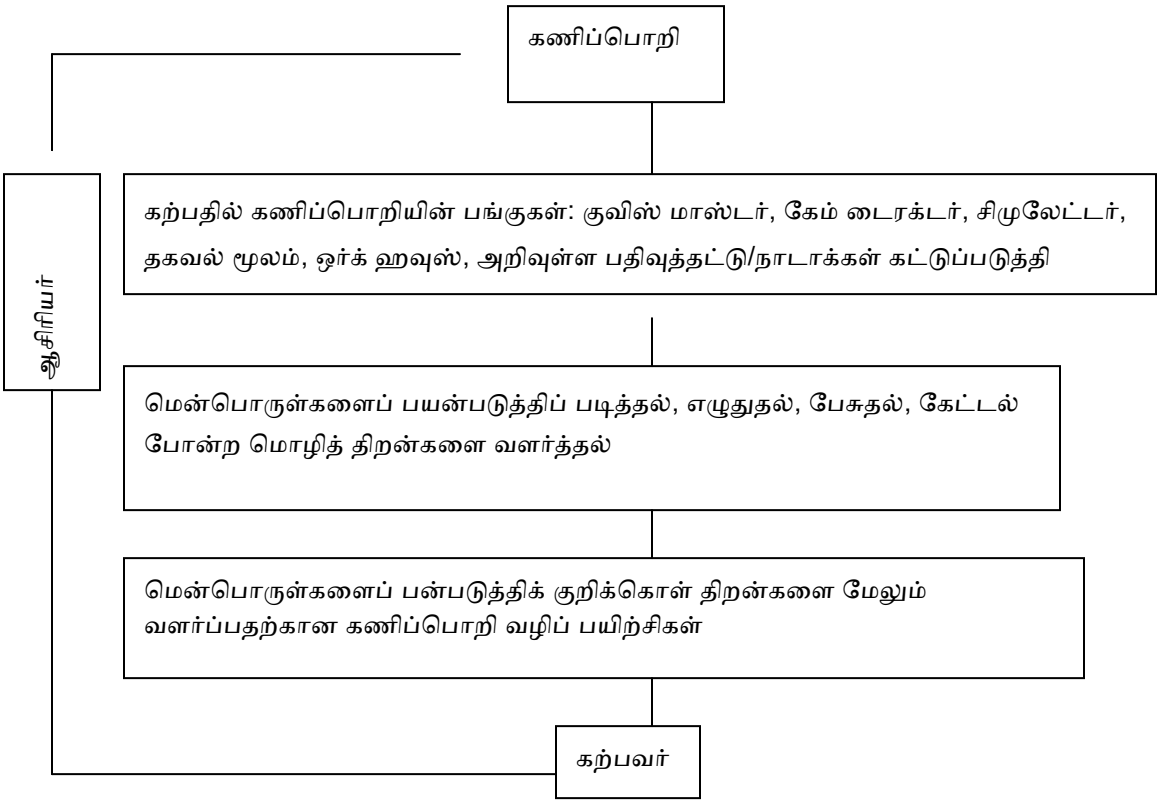
இயந்திர மொழிபெயர்ப்புக்குப் பொருண்மையியல் ஆய்வி மற்றும் ஆக்கி, தொடரியல் ஆய்வி மற்றும் ஆக்கி, மொழிபெயர்ப்பு அகராதி என்பன போன்ற துணைக்கருவிகளைத் தயாரித்தல் வேண்டும்.

மொழிகற்றல்/கற்பித்தல்

மொழியைக் கற்பதற்கும் கற்பிப்பதற்கும் கணிப்பொறியைப் பயன்படுத்தலாம். வெறும் எண், எழுத்து இவற்றை மட்டும் கையாண்டு வந்த கணிப்பொறி எண், எழுத்து, படம் (Graphics), ஒலி (Audio), ஒளி (Video) அசையும் உருவங்கள் (Animation) எனப்பலவற்றை ஏற்றுக்கொண்டுள்ளது. இப்படி வேறுபட்ட பன்முகத் தொடர்புகள் கொண்ட ஒருங்கிணைப்பு மல்டிமீடியா (Multimedia) எனப்படும். மல்டிமீடியாவைப் பயன்படுத்தித் தமிழ்க்கற்றலையும் கற்பித்தலையும் ஆர்வத்தைத் தூண்டும் செயலாக மாற்றலாம்.

கணிப்பொறி அடிப்படையிலான கணிப்புச் செயல்களைக் கல்வி மற்றும் உளவியல் அடிப்படையிலான கற்றலுக்கும் கற்பித்தலுக்கும் பயன்படும் படி செய்வதுதான் கணிப்பொறி மூலம் கற்றல்/கற்பித்தல் மென்பொருள்களைத் தயாரித்தலின் நோக்கமாகும். இங்குக் கணிப்பொறி, கற்றல் அடிப்படையிலான மொழிச்செயல்பாடுகளைச் செய்வதற்குப் பல பங்குகள் வகிக்கும்.

எடுத்துக்காட்டாக குவிஸ்மாஸ்டர், கேம் டைரக்டர், சிமுலேட்டர், ஒர்க் ஹவுஸ், தகவல் மூலம் கட்டுப்படுத்தி போன்ற பல பங்குகளைக் கணிப்பொறி


கணிப்பொறி
ஆசிரியர்
கற்பதில் கணிப்பொறியின் பங்குகள்: குவிஸ் மாஸ்டர், கேம் டைரக்டர், சிமுலேட்டர், தகவல் மூலம், ஒர்க் ஹவுஸ், அறிவுள்ள பதிவுத்தட்டு/நாடாக்கள் கட்டுப்படுத்தி
மென்பொருள்களைப் பயன்படுத்திப் படித்தல், எழுதுதல், பேசுதல், கேட்டல் போன்ற மொழித் திறன்களை வளர்த்தல்
மென்பொருள்களைப் பன்படுத்திக் குறிக்கொள் திறன்களை மேலும் வளர்ப்பதற்கான கணிப்பொறி வழிப் பயிற்சிகள்
கற்பவர்

வகிக்கும். பின்வரும் வரைபடம் கணிப்பொறியின் உதவியுடன் கூடிய கற்றல்/கற்பித்தல் செயற்பாட்டைத் தெளிவுபடுத்தும்.

மேற்கண்டவாறு கணிப்பொறியைப் பயன்படுத்தி மொழித்தொடர்பான செயற்பாடுகளைச் செய்ய மொழியியலாரும் கணிப்பொறியியலாரும் இணைந்து செயலாற்ற வேண்டும். ஒருவருடைய செயற்பாட்டை மற்றவர் புரிந்துகொண்டு செயலாற்றினால் மொழித்தொடர்பான கணிப்பொறிக் கருவிகளை, மென்பொருள்களை எளிதில் உருவாக்கிப் பயன்படுத்தலாம்.

இயல் 2

எழுத்துப்பிழை திருத்தமும் இலக்கணப்பிழை திருத்தமும்

2.0 முன்னுரை

சொல்லாய்வியின் வழி ஆவணங்கள் (Documents) உருவாக்கப் படுகின்றன. ஆவணம் ஒரு கடிதமாகவோ சிறுகுறிப்பாகவோ கட்டுரையாகவோ நூலாகவோ இருக்கலாம். சொல்லாய்வியில் ஆவணத்தை உருவாக்கப் பொதுவாகப் பின்வரும் மூன்று முறைகள் பயன்படுத்தப்படுகின்றன:

1. கணிப்பொறியின் விசைப்பலகை (Keyboard) வழி உள்ளீடு செய்தல்
2. பேச்சுணரி ஒழுங்கு முறையைப் (Voice Recognition System (VRS)) பயன்படுத்தி உரையை உள்ளீடு செய்தல்: இவ்வாறு செய்யும் பொழுது ஒலிவடிவில் பேசப்படும் செய்திகள் உரைவடிவிற்கு மாற்றப்படுகின்றன. இதற்குப் பேச்சிலிருந்து உரை (Speech to Text (STT)) என்ற தொழில் நுட்பம் உதவுகிறது.
3. தட்டச்சு செய்தல் அல்லது அச்சடிக்கப்பட்ட உரையை ஒளிவழி எழுத்துணர்தல் (Optical Character Recognition (ORC)) என்ற தொழில்நுட்பத்தின் மூலம் களிப்பொறிக்கு உள்ளீடு செய்து சொல்லாய்விற்கு ஆவணமாக மாற்றுவது.

உரையை உள்ளீடு செய்ய மேற்கண்ட மூன்று முறைகளுள் ஏதேனும் ஒன்றைப் பின்பற்றலாம். பின்னர் சொல்லாய்வியின் மூலம் அவைகளைக் கையாளலாம். உரைக்குள் எழுத்துப் பிழைகளும் இலக்கணப் பிழைகளும் இருப்பதற்கான வாய்ப்புகள் உள்ளன. ஆங்கில உரைகளில் உள்ள எழுத்துப்

பிழைகளைக் கணிப்பொறிதானாகவே புரிந்துக்கொண்டு மாற்றுவதற்குச் சொல்லாய்வியில் வசதிகள் உள்ளன. இதற்கு எழுத்துப் பிழைகளைக் கண்டறிந்து திருத்துதல் என்று பெயர். இத்தகைய பிழைத்திருத்தி தமிழ் சொல்லாய்விகளில் உருவாக்க முயற்சிகள் எடுக்கப்பட்டு வருகின்றன. எழுத்துப் பிழைகள் நிகழ்வதற்கான வாய்ப்புகள், எழுத்துப் பிழைகளின் வகைகள், பிழைகளைக் கண்டறிந்து திருத்துவதற்கான செயல்பாட்டு வழிமுறைகள் போன்றவையெல்லாம் விரிவாக ஆராய்ந்து செயல்படவேண்டி உள்ளது.

பிழைகளை எழுத்துப்பிழைகள், இலக்கணப்பிழைகள் என்று வரையறுப்பது கவனக்குவிப்பு செய்வதற்கு எளிமையாக அமைந்தாலும் இரண்டையும் வெவேறான செயல்பாடுகள் என்று கூறுவது கடினம். எழுத்துப்பிழைகள் இலக்கணப்பிழைகளுக்கு வழிவகுக்கலாம் அல்லது இலக்கணப்பிழைகள் எழுத்துப்பிழைகளாகத் தோன்றலாம். குறிப்பாகத் தமிழ் போன்ற திரிபு ஒட்டுகள் மிகுந்த மொழியில் எழுத்துப்பிழை இலக்கணப்பிழைக்கும் இலக்கணப்பிழை எழுத்துப்பிழைக்கும் வழிவகுக்கும். சந்திப் பிழைகளை அல்லது புணர்ச்சிப் பிழைகளை எழுத்துப்பிழைகளாகவும் இலக்கணப்பிழைகளாகவும் கருதலாம்.

2.1. எழுத்துப்பிழை திருத்தம்

எழுத்துப்பிழை திருத்தம் எழுத்துப்பிழை நிகழ்வதற்கான வாய்ப்புகளைக் கண்டறிதல், எழுத்துப்பிழைகளக் கண்டறிதல், பிழை நீக்கும் வழிமுறைகளைக் கண்டறிதல், பிழை நீக்குவதில் உள்ள சிக்கல்களை அறிதல் என்பன போன்ற செயல்பாடுகளை உட்படுத்தும்.

2.1.1 எழுத்துபிழைகள் நிகழ்வதற்கான வாய்ப்புகள்

உரைச் செய்திகளைக் கீழ்கண்ட ஏதேனும் ஒருங்குமுறைகளில் உள்ளீடு செய்யப்படும்போது எழுத்துப்பிழைகள் நிகழ்வதற்கு வாய்ப்புகள் உள்ளன.

- விசைப்பலகை வழி உள்ளீடு செய்யும்பொழுது தட்டச்சுப் பிழைகள் (Typographical Errors) ஏற்படுகின்றன.
- பேச்சுணரி ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்றபொழுது சொல்லின் உச்சரிப்பு மாற்றங்களினால் பிழைகள் ஏற்படுகின்றன. இத்தகைய பிழைகளுக்கு உச்சரிப்பு வேறுபாடு (Pronunciation Variation) என்று பெயர்.
- ஒளி எழுத்துணரி (ORC) வழி உள்ளீடு செய்கின்றபொழுது கருவிகளின் குறைபாடுகளினாலோ அல்லது எழுத்துக்களைச் சரியாகக் கருவிகள் படித்தறிய இயலாததாலோ பிழைகள் தோன்றுகின்றன. எடுத்துக்காட்டாக ஆங்கிலத்தில் l என்ற எழுத்திற்கும் 1 என்ற எழுத்திற்கும் வேறுபாடு தெரியாததால் பிழை ஏற்படுகின்றன. பின்வரும் அட்டவணையில் சில சரியான எழுத்துக்களும் அவற்றிற்கிணையான தவறாகப் புரிந்துகொள்ளப்படும் எழுத்துக்களும் பட்டியலிடப்பட்டுள்ளன:

சரியான எழுத்து	தவறாகப் புரிந்துகொள்ளப்பட்ட எழுத்து
l	1
D	0
ri	n
0	θ

மேற்கண்டவாறு மூன்றுவகையான உள்ளீட்டு முறைகளிலும் அதற்கான கருவிகளைப் பயன்படுத்தும் பொழுது பிழைகள் தோன்றுகின்றன. பேச்சறிதல் ஒழுங்குமுறை மூலம் உள்ளீடு செய்கின்றபொழுது ஒரே வார்த்தையை அதன் இலக்கணக் கூறு மற்றும் தன்மைகளுக்கு ஏற்ப வெவ்வேறு விதமாக உச்சரிக்க வேண்டியுள்ளது. எடுத்துக்காட்டாக ஆங்கிலத்தில் Object என்ற சொல் பெயர்ச் சொல்லாகவும் வினைச் சொல்லாகவும் பயன்படுகின்றது. பெயர்ச்சொல்லாக பயன்படும்போது 'ஆப்ஜிக்ட்' என்றும் வினைச் சொல்லாகப் பயன்படுத்தும்போது 'அப்ஜெக்ட்' என்றும் உச்சரிப்பு செய்யவேண்டியுள்ளது. அதைப்போல் The என்ற சொல்லைச் சூழலுக்கு ஏற்ப [ðə] என்றும் என்றும் [ði] மாறுப்பட்ட இருவகைகளில் உச்சரிக்க வேண்டியுள்ளது. இவ்வாறாக, ஒரே சொல்லிற்கான உச்சரிப்பு மாற்றங்களுக்குத் திட்டவட்டமாக நன்கு வரையறுக்கப்பட்ட விதிமுறைகள் உள்ளன. அவ்விதி முறைகளைக் கணிப்பொறிக்கு ஏற்ற வண்ணம் வடிவமைப்புகளாக வடிவமைத்து இத்தகைய பிழைகள் நீக்கப்படுகின்றன.

2.1.2 எழுத்துப் பிழைகளைக் கண்டறிதல் – சரிசெய்தல் சிக்கல்களின் வகைகள்

குகிச் (Kukich, 1992) தமது மதிப்பீட்டாய்வு கட்டுரையில் எழுத்துப்பிழை திருத்தத்தை பரந்த அளவில் மூன்றாகப் பிரிகின்றார். இதன் படி எழுத்துப்பிழைகளைக் கண்டறிதல்-சரிசெய்தல் தொடர்பான சிக்கல்கள் பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

- மொழியில் இல்லாத சொற்களைக் கண்டறிதல் (Non-Word Error Detection)
- எழுத்துப்பிழை உள்ள செய்திகளை மட்டும் தனிமைப்படுத்தி ஆராய்தல் (Isolated-word error correction)
- எழுத்துப்பிழை உள்ள சொல்லின் முன்பின் உள்ள அல்லது அச்சொல் இடம்பெற்ற தொடரின் பொருளையமாகக் கொண்டு பிழைநீக்கி சரிசெய்தல் (context-dependent error detection and correction)

குகிச் (Kukich 1992) மனிதர்களின் தட்டச்சுப் பிழைகளை இரு பிரிவுகளாகப் பிரிக்கிறார்:

1. தட்டச்சுப் பிழைகள் (typographic errors)
2. புரிதல் பிழைகள் (cognitive errors) சொல்லின் எழுத்து வடிவம் உரையாளருக்குத் தெரியாததால் ஏற்படும் பிழைகள்

டமரு (Dameraru, 1964) தவறாக எழுதப்பட்ட சொற்களில் 80% தனிநிலைத் தவறால் ஏற்பட்ட பிழைகளாகக் காண்கிறார். இதன் படி தட்டச்சுப் பிழைகளைப் பின்வரும் நான்கு வகைகளாகப் பிரிக்கலாம்:

1. விடுபடு பிழை (Deletion Error): இப்பிழையில் ஒரு எழுத்து விடுபட்டிருக்கும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் *'ட்'* என்ற எழுத்தில்லாமல் *'சடம்'* என இருப்பது விடுபடு பிழையாகும்.
2. செருகல் பிழை (Insertion Error): இப்பிழையில் ஒரு எழுத்து கூடுதலாக இருக்கும். எடுத்துக்காட்டாக, *சட்டடம்* என்ற பிழையான சொல்லில் *'ட'* என்ற எழுத்துக் கூடுதலாக இருப்பது செருகல் பிழையாகும்.

3. பதிலீட்டுப் பிழை (Substitution Error): இப்பிழையில் ஒரு எழுத்திற்குப் பதிலாக வேறொரு எழுத்தை மாற்றி அமைத்து விடும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் முதல் எழுத்து *'ச'* தவறாக அடிக்கப்படும்போது அத்தவறான எழுத்து 'ச' ஆகவோ 'க' ஆகவோ 'ப' ஆகவோ இருக்கலாம். இம்மாதிரியான பிழைகளைப் பதிலீட்டுப் பிழைகள் எனலாம்.
4. இடமாற்றுப் பிழை (Transposition Error): இப்பிழையில் அடுத்தடுத்த இரண்டு எழுத்துக்கள் இடம் மாற்றி இருக்கும். எடுத்துக்காட்டாக, *சட்டம்* என்ற சொல்லில் பிழையாக *'ட'* என்ற எழுத்தும் *'ட்'* என்ற எழுத்தும் இடம் மாறி *சடட்ம்* என்று மாற்றி அடிக்கப்படுவது இடமாற்றுப் பிழையாகும்.

2.1.3 மொழியில் இல்லாத சொற்களைக் கண்டறிந்து மாற்றுதலுக்கான செயல் திட்டம்

ஒரு மொழியில் உள்ள சொல்லினைக் கண்டறிந்து சரிசெய்வதற்கு அகராதி பயன்படுத்தப்படுகிறது. கொடுக்கப்பட்ட சொல் அகராதியில் இல்லை எனில் அச்சொல்லை மொழியில் இல்லாத சொல் எனக் கருதலாம். இதற்குத் தகுந்த கணிப்பொறிக்கான அகராதி சிறியதாக இருத்தல் நலம். மிகப் பெரிய அகராதியாக இருந்தால் அதிகமாகப் புழக்கத்தில் உள்ள எழுத்துப்பிழைச் சொற்கள் கூட அரிதாகப் பயன்படுத்தும் சரியான சொல்லாக இருக்கும்.

2.1.4 பிழை நீக்கும் வழிமுறைகள்

உரைச் செய்திகளில் உள்ள எழுத்துப் பிழைகளைக் கண்டறிந்து மாற்றுவதற்கு இருவழி முறைகள் பயன்படுத்தபடுகின்றன:

1. பெய்சியன் விதிகள் (Bayesian Rules)
2. ஓசை ஊடக மாதிரி (Noisy Channel Model)

பெய்சியன் விதிமுறைகளைப் பின்பற்றும் பொழுது எழுத்துப் பிழைகளைக் கண்டறிவதற்கான வாய்ப்புகளை ஆராய்வதற்குக் கணித முறையில் அமைந்த நிகழ்தகவுக் கோட்பாடு (Probability Theory) பயன்படுத்தப்படுகிறது. பெறும்பாலும் இம்முறைகள் தட்டச்சு செய்த உரைகளுக்குப் பயன்படுத்தப்படுகின்றன. இம்முறையை ஒளிவழி எழுத்துணரி (OCR) முறையிலும் பேச்சுப் புரிந்துகொள்ளல் ஒழுங்குமுறைகளிலும் (Voice Recognition System) உள்ளீடு செய்யப்படும் உரைகளுக்கும் பயன்படுத்தலாம்.

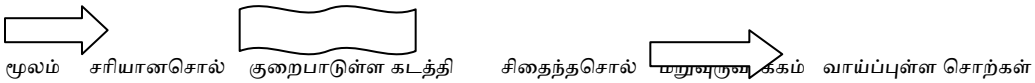
இரண்டாவது ஓசை ஊடக மாதிரி (Noisy Channel Model) முறையில் உள்ளீடு செய்யப்படுகின்ற உரைகள் எவ்வாறு மாற்றமடைகின்றன, மாற்றமடைவதற்கான காரணங்கள் யாவை போன்றவைகள் ஆராயப்பட்டு ஒரு மாதிரி வடிவம் (Model) தயாரிக்கப்படுகின்றது. பின்னர் இம்மாதிரி வடிவத்திற்கான எதிர்மறைச் செயல்பாடுகள் (Reverse Process) ஒருமாதிரிப் படிவமாக உருவாக்கப்பட்டுப் பிழைகள் நீக்கப்படுகின்றன. எடுத்துக்காட்டாக, தொலைபேசியில் பேசுகின்றபொழுதும் ஒலிவாங்கி வழி பேசுகின்றபொழுதும் ஒருவர் பேசும் குரல் அமைப்புகள் மாறுபடுகின்றன. இத்தகைய கருவிகள் மூலம்

கிடைக்கப்பெறுகின்ற முடிவுகள் ஒலி அமைப்பைப் பொறுத்துள்ளது; உள்ளபடியாகவே அமையாது. இந்நிலையில் அம்மாறுபாடுகள் நிகழ்வதற்கான காரணங்களை ஆராய்ந்து அதற்கு எதிர்மறையான திட்டங்களை வடிவமைக்கின்ற பொழுது இவைகள் மூலம் நிகழும் பிழைகளை நீக்கலாம்.

2.1.4.1 பெசியன் முறை (Bayesian Medhod)

ஒரு உரையிலுள்ள எழுத்துப்பிழைகளைக் கண்டறிந்து திருத்துவதற்குப் பல செயல் திட்ட முறைகள் வடிவமைக்கப்படுகின்றன. அவற்றில் பெசியன் முறை நிகழ்தகவு அடிப்படையில் (Probability Thories) அமைந்ததாகும்.

இம்முறையின் அடிப்படைத் தத்துவம் பின்வரும் படத்தில் விளக்கப்பட்டுள்ளது:



உரைக்கான சொற்கள் தகுந்த உள்ளீட்டுக் கருவிகள் மூலம் உள்ளீடு செய்யப்படுகின்றன. உள்ளீட்டுக் கருவி கணிப்பொறியின் விசைப்பலகையாகவோ ஒலியுணரி ஒழுங்குமுறையின் ஒலிவாங்கியாகவோ இருக்கலாம். தட்டச்சு செய்யும்பொழுது எழுத்துக்களின் சிதைவோ ஒலியுணரி ஒழுங்குமுறையில் ஒலிச் செய்தியின் சிதைவோ ஏற்பட வாய்ப்புகள் உள்ளன. மூலக்கருவிகள் மூலம் உருவாக்கப்படும் சொல் உரிய கடத்திகள் (Channels) மூலம் கடத்தப்படுகிறது. கடத்திகளில் உள்ள குறைபாடுகள் காரணமாகவும் கடத்தும் செய்திகளில் சிதைவு ஏற்படலாம். இவ்வாறாகச் சிதைந்த சொல் மறு உருவாக்கம் (Decode) செய்யப்படுகிறது. இச்சொல் மூலக்கருவி மூலம் உருவாக்கப்பட்ட சொல்லுக்கு வாய்ப்புள்ள சொல்லாகவே அமையும். இவ்வாறு சிதைந்த சொல்லுக்கு வாய்ப்புள்ள சொற்களின் பட்டியல் உருவாக்கப்படும்.

இப்பட்டியலிலிருந்து சிதைந்த சொற்கூறு திருத்தம் செய்யப்பெற்றுப் பிழை நீக்கிய சரியான சொல் அடையாளம் காணப்படுகிறது.

பேசியன் முறையில் எழுத்துத் திருத்திக்கான வழிமுறைகள் இருபெரும் கட்டங்களாகச் செய்யப்படுகின்றன:

1. பிழை நீங்கிச் சரியான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் தயாரித்தல்

2. இச்சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லுக்கும் மதிப்பெண் அளித்து அதிக மதிப்பெண் பெறும் சொல்லைப் பிழைத் திருத்தம் செய்யப்பட்ட சொல்லாகக் கருத வழிவகை செய்தல்.

இம்முறையில் பிழைத்திருத்தம் என்பது சொல்லில் ஏதேனும் ஒரு எழுத்து மட்டும் பிழையாக இருக்கலாம் என்ற வாதத்தின் அடிப்படையில் செயல்முறைகள் மேற்கொள்ளப்படுகின்றன. எனவே பின்வரும் வகையிலான எழுத்து பிழைகள் மட்டும் இம்முறையில் எடுத்துக் கொள்ளப்படுகின்றன:

- சொல்லின் ஒரு எழுத்து விடுபட்டு இருத்தல்
- சொல்லில் ஒரு எழுத்து அதிகமாக இருத்தல்
- சொல்லில் ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து தவறாக இருத்தல்
- சொல்லில் அடுத்தடுத்த இரு எழுத்துக்கள் இடம் மாறி இருத்தல்

பேசியன் முறையில் எழுத்துத் திருத்தி உருவாக்குவதற்கு முதலில் பெருந்தரவு (Corpus) ஒன்றினைக் கணிப்பொறியில் உருவாக்கிக்கொள்ள வேண்டும். இந்தப் பெருந்தரவில் மொத்தம் ‘N’ சொற்கள் இருப்பதாகக் கொள்வோம். பிழையானச் சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியல் ஒன்றை உருவாக்க வேண்டும். இப்பட்டியலை உருவாக்கும் ஒரு சொல்லில் ஒரு எழுத்தினால் மட்டும் பிழை நிகழ்வதாக எடுத்துக்கொள்ளப்படுகிறது. எனவே சொல்லில் (அ) ஒரு எழுத்து

விடுபட்டு இருத்தல், (ஆ) ஒரு எழுத்து கூடுதலாக இருத்தல், (இ) ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருத்தல், (ஈ) அடுத்தடுத்த இரு எழுத்துக்கள் தவறாக இடம் மாறியிருத்தல் ஆகிய நான்கு மட்டும் கருத்தில் கொள்ளப்படுகின்றன. மேற்கண்ட முறைகளின்படி பிழையான சொல்லிற்கு வாய்ப்புள்ள சொற்பட்டியலில் 'V' சொற்கள் இருப்பதாகக் கொள்வோம். இப்பட்டியலில் 'c' என்ற சொல்லாக இருப்பதற்கான முன்நிகழ்வு எண்ணிக்கை பின்வரும் சூத்திரத்தின் மூலம் கணக்கிடப்படுகிறது:

$$P(c) = \frac{C(c)+0.5}{N+0.5V}$$

(V = வாய்ப்புள்ள சொல், C = விரிதரவு 'Corpus')

மேற்கண்ட சூத்திரம் ஒரு சொல் மட்டும் பெருந்தரவில் இடம்பெற்றுள்ள முன்நிகழ்வு (Prior) மதிப்பைத் தருகிறது. இரண்டாவது கட்டநிலையில் தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்களுக்கு மதிப்புகள் கண்டறியப்படுகின்றன. இதற்கு மாதிரி உரைகள் ஒரு தட்டச்சர் அல்லது பல தட்டச்சர்களைக் கொண்டு உள்ளீடு செய்யப்படுகின்றன. கெர்னிகன் மற்றும் பிறர் (Kernighan et al) ஒவ்வொரு வகைத் தனிநிலைப் பிழைக்கும் நான்கு குழப்ப அணிகளைப் (confusion matrices) பயன்படுத்தினர்.

1. Del (x, y) என்பது xy என்ற இரு எழுத்துகளுக்குப் பதிலாக x மட்டும் எத்தனை முறை தவாற அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.

2. Ins (x, y) என்பது x என்ற எழுத்துக்களுக்குப் பதிலாக xy என்று தட்டச்சு செய்துள்ள எண்ணிக்கையைக் குறிக்கிறது.

3. Sub (x, y) என்பது x என்ற எழுத்திற்குப் பதிலாக எத்தனை முறை y என்ற எழுத்து தட்டச்சு செய்யப்பட்டுள்ளது என்பதைக் குறிக்கிறது.

4. Trans (x, y) என்பது xy என்ற எழுத்துக்களுக்குப் பதிலாக yx என இடமாறி அச்சடிக்கப்பட்டுள்ளது என்பதைக் குறிக்கிறது.

தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் இருந்து ஏதாவது ஒரு சொல் 'c' என்பதை எடுத்துக்கொள்வோம். 'c_p' என்பது 'c' என்ற சொல்லில் P- இடத்தில் உள்ள எழுத்தைக் குறிப்பதாகக் கொள்வோம். இந்நிலையி 'c' என்ற சொல் வாய்ப்பான சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

1. ஒரு எழுத்து விடப்பட்டிருக்கும் நிலையில்

$$P(t/c) = \frac{del[c_p - 1, c_p]}{count\ [c_{p-1} c_p]}$$

2. ஒரு எழுத்து கூடுதலாக இருக்கும் நிலையில்

$$P(t/c) = \frac{ins\ [c_{p-1}, t_p]}{count\ [c_{p-1}]}$$

3. ஒரு எழுத்திற்குப் பதிலாக மற்றொரு எழுத்து இருக்கும் நிலையில்

$$P(t/c) = \frac{trans\ [t_p, c_p]}{count\ [c_p]}$$

4. இரு எழுத்துக்கள் இடம் மாறி இருக்கும் நிலையில்

$$P(t/c) = \frac{trans[c_p, c_{p+1}]}{count\ [c_p, c_{p+1}]}$$

தவறான சொல் வாய்ப்புள்ள சொற்பட்டியலில் உள்ள 'c' என்ற சொல்லாக இருப்பதற்கான மதிப்பு பின்வருமாறு கணக்கிடப்படுகிறது:

$$W= P (t/c)*P(c)$$

இந்த சூத்திரத்தில் P(c) என்பது c என்ற சொல் பெருந்தரவில் உள்ள முன் நிகழ்வு மதிப்பைத் தருகிறது. P(t/c) என்பது தவறாகத் தட்டச்சு செய்யப்பட்ட சொல் 'c' என்ற சொல்லாக இருப்பதற்கு எவ்வளவு வாய்ப்புகள் உள்ளன என்ற மதிப்பைத் தருகின்றது. இவ்விரு எண்களின் பெருக்கல் தொகை, அதாவது P (t/c)*P(c) என்பது இச்சொல்லிற்கான மதிப்பெண்ணைத் தருகிறது.

மேற்கண்டவாறு தவறான சொல்லுக்கு வாய்ப்புள்ள சொற்பட்டியலில் உள்ள ஒவ்வொரு சொல்லிற்கும் மேற்கண்ட மதிப்பினைக் கண்டறிய வேண்டும். எந்தச் சொல் அதிக மதிப்பெண் பெற்றிருக்கிறதோ அந்தச் சொல் மிகச் சரியான சொல் என கருதப்படுகிறது. அதிக மதிப்பெண் பெற்றிருக்கும் சொல் பின்வரும் சூத்திரத்தின் மூலம் கூறப்பட்டுள்ளது.

$$W= \underset{c \in V}{\operatorname{argmax}} \frac{P(t/c)P(c)}{P(t)}$$

மேற்கண்டவாறு அதிக மதிப்பெண் பெற்ற சொல் இடம் பெற்ற வாக்கியத்தின் பொருளுடன் ஒப்பிடுப் பார்க்கப்படும். அது வார்த்தையின் பொருளுக்கேற்ற சொல் இல்லையெனில் அடுத்த மதிப்பெண் பெற்ற சொல் பொருத்திப்பார்க்கப்படும். இவ்வாறாகத் தவறாகத் தட்டச்சு செய்யப்பட்டுள்ள சொற்களைச் சரியான சொற்களாக மாற்றச் செயல் திட்டங்கள் உள்ளன.

இவ்வாறாக உரைகள் எழுதும்போது பிழைகளைக் கண்டறிந்து திருத்துவதற்குக் கணிப்பொறியில் பல செயல்திட்டங்கள் உருவாக்கப்படுகின்றன. அதில் மிகவும் பிரபலமானது மேற்கண்ட பேசியன் முறையாகும். இதுபோல் குறைந்த திருத்த தூரம் (minimum edit distance) என்ற மற்றொரு முறையும் பயன்படுகிறது.

2.1.4.2. குறைந்த திருத்து தூரம்

முந்தைய பகுதியில் பேசியன் வழிமுறைவரைவு எவ்வாறு திருத்தங்களைச் செய்ய முடிந்தது என்று பார்த்தோம். இருப்பினும் கெர்னிகன் மற்றும் பிறர் (Kernighan et al 1990) ஒரு சொல் ஒரு எழுத்துப்பிழை மட்டும் கொண்டிருக்கும் என்ற எளிய ஊகத்தைச் சார்ந்திருந்தனர். நாம் பன்முகப் பிழைகளைக் கையாளுவதற்குக் கூடுதல் ஆற்றலுள்ள வழிமுறை வரைவை விருப்பலாம். அம்மாதிரியான வழிமுறைவரைவைக் கோர்வை தூரத்தின் சிக்கலுக்குரிய பொதுவான தீர்வாகக் கருதலாம். 'கோர்வை தூரம்' என்பது எந்த அளவுக்கு இரண்டு கோர்வைகள் ஒன்றுபோல் இருக்கின்றன என்பதன் ஒரு அணியாகும். பேசியன் முறையைப் பிழைதிருத்தச் சிக்கலுக்கு அம்மாதிரியான வழிமுறைவரைவைப் பயன்படுத்தும் வழியாகப் பார்க்க இயலும் ; நாம் தரப்பட்ட பிழைக்கு மிக உயர்ந்த நிகழ்தகவு கொண்டது என்ற அர்த்தத்தில் தவறுக்கு மிக அண்மையில் உள்ள சொல்லைத் தேர்வுச் சொல்லாக எடுக்கின்றோம்.

கோர்வை தூரத்தைக் கண்டுபிடிப்பதற்கு வழிமுறைவரைவுகளின் பிரபலமான வகுப்புகளில் ஒன்று வாக்னர் மற்றும் ஃபிஷர் (Wagner and Fischer 1974) என்பவர்களால் குறைந்த திருத்து தூரம் (minimum edit distance) என்று பெயரிடப்பட்ட (ஆனால் தனித்தனியாகப் பலரால் கண்டுபிடிக்கப்பட) ஒரு

பதிப்பாகும். இரு கோர்வைகளுக்கு இடையே உள்ள குறைந்த திருத்து தூரம் என்பது ஒரு கோர்வையை மற்றொரு கோர்வையாக மாற்றத் தேவைப்படும் திருத்து நடவடிக்கைகளின்/செயன்மைகளின் (செருகல், நீக்குதல், பதிலீடுசெய்தல்) குறைந்த எண்ணிக்கையாகும். எடுத்துக்காட்டாக intension மற்றும் extension என்ற சொற்களுக்கு இடையே உள்ள இடைவெளி ஐந்து செயன்மைகளாகும்/நடவைக்கைகளாகும்; இதை மூன்று வழிகளில் உருப்படுத்தம் செய்யவியலும் ; கீழே காட்டியுள்ளது போன்ற சுவடு/தடம் (trace), வரிசையாக்கம் (alignment) அல்லது செயன்மைப் பட்டியல் (operation list) என்ற மூன்று வழிகள் உள்ளன.

சுவடு

I n t e n t i o n
e x e c u t i o n

வரிசையாக்கம்

I n t e n ε t i o n
ε e x e c u t i o n

நடவடிக்கைப் பட்டியல்

I n t e n t i o n

l-ஐ நீக்கவும்

n t e n t i o n

n-ஐ e-ஆல் பதிலீடு செய்யவும்.

e t e n t i o n

t-ஐ x-ஆல் பதிலீடு செய்யவும்.

e x e n t i o n

u-ஐ செருகவும்

e x e n u t i o n

n-ஐ c-ஆல் பதிலீடு செய்யவும்

e x e c u t i o n

நாம் ஒவ்வொரு நடவடிக்கைக்கும் ஒரு குறிப்பிட்ட மதிப்பையோ (cost) சுமையையோ (weight) ஒதுக்கலாம். இரு வரிசைகளுக்கு இடையிலான லெவென்ஷ்டைன் தூரம் தான் மிக எளிய அளவிடும் காரணி; இதில் ஒவ்வொரு நடவடிக்கைக்கும் விலை 1 (Levenshtein 1966) ஆகும். இதன்படி intension என்பதற்கும் execution என்பதற்கும் இடையில் உள்ள தூரம் 5 ஆகும். லெவென்ஷ்டைன் தன்னுடைய அளவுமுறைக்கு மற்றொரு மாற்று பதிப்பை முன்மொழிகின்றார்; இதில் ஒவ்வொரு செருக்கலுக்கும் அல்லது நீக்கலுக்கும் மதிப்பு ஒன்று ஆகும்; பதிலீடு அனுமதிக்கப்படமாட்டாது (பதிலீட்டை ஒரு செருகலாலும் ஒரு நீக்கலாலும் உருப்படுத்தம் செய்ய இயலும் ஆகையால் ஒவ்வொரு பதிலீட்டுக்கும் மதிப்பு 2 தரப்படுகின்றது). இந்தப் பதிப்பின் பயன்பாட்டின் படி intension என்பதற்கும் execution என்பதற்கும் இடையிலுள்ள தூரம் 8 ஆகும். நாம் கூடுதல் கலவைத்தன்மையான செயல்பாடுகளால் சுமையேற்றலாம்; எடுத்துக்காட்டாக ஒவ்வொரு நடவடிக்கைக்கும் முன்னர் கூறிய குழப்ப அணிகளைப் பயன்படுத்தி ஒரு நிகழ்தகவை தருவது. இந்நேர்வில் இரண்டு கோர்வைகளுக்கு இடையிலுள்ள குறைந்த திருத்து தூரத்தைப் பற்றி பேசுவதற்குப் பதிலாக நாம் ஒரு கோர்வைக்கும் மற்றொரு கோர்வைக்கும் அதிக நிகழ்தகவு வரிசையாக்கம் பற்றி பேசுவோம். இதை நாம் செய்தால் ஒவ்வொரு மாற்றத்தின் நிகழ்தகவுகளைப் பெருக்கும் பெரிதாக்கப்பட்ட குறைந்த திருத்து தூர வழிமுறைவரைவை ஒரு தேர்வுப் பிழை தரப்படுகையில் பல் தவறு தட்டச்சின் பேசியன் சாத்தியத்தியத்தை மதிப்பிட பயன்படுத்த இயலும்.

ஆற்றலுள்ள வழிமுறை அமைப்பால் (dynamic programming) குறைந்த திருத்து தூரம் கணக்கிடப்படுகின்றது. ஆற்றலுள்ள வழிமுறை அமைப்பு வழியமைப்பு வரைவுகளின் ஒரு வகுப்பின் பெயராகும்; இது முதலில் பெல்மான் (Bellman 1957) என்பவரால் அறிமுகப்படுத்தப்பட்டது; இது துணைச் சிக்கல்களின் திர்வுகளை ஒன்றிணைத்து சிக்கல்களைத் தீர்க்க அட்டவணை-இயக்க முறையைப் (table-driven method) பயன்படுத்துகின்றது. இந்த வழியமைப்பு வரைவுகளின் வகுப்பு பேச்சு மற்றும் மொழி ஆய்வுகளில் கூடுதல் பொதுவாகப் பயன்படுத்தப்படுகின்ற வழியமைப்பு வரைவுகளை உட்படுத்தும்; எழுத்துப் பிழை திருத்ததிற்கான குறைந்த திருத்து தூர வழியமைப்பு வரைவு, பேச்சு புரிந்துகொள்வதிலும் இயந்திர மொழிபெயர்ப்பிலும் பயன்படுத்தப்படும் வெட்டர்பி வழியமைப்பு வரைவு (Veterbi algorithm) மற்றும் பகுத்துக்குறிக்கப் பயன்பத்தப்படும் CYK மற்றும் Early வழியமைப்பு வரைவு என்பன இவையாகும்.

2.2 இலக்கணப்பிழை திருத்தம்

இலக்கணத் திருத்தி ஒரு வாக்கியத்தின் தொடரியல் செம்மையை நிர்ணயிக்கும் அல்லது தீர்மானிக்கு இலக்கணத் திருத்தம் சொல் ஆய்விகளிலும் (word processor) தொகுப்பிகளிலும் (compilers) பெரும்பாலும் பயன்படுத்தப்படுகின்றது. தொகுப்பி போன்ற பயன்பாடுகளில் இலக்கணத் திருத்தம் எளிமையானதாகும். எனென்றால் சொற்றொகை அவற்றில் எல்லைக்குப்பட்டதாகும்; ஆனால் இயற்கை மொழிகளில் இது சவாலுக்குரியதாகும்; ஏனென்றால் அதன் சொற்றொகை எல்லைக்கு உட்பட்டதல்ல.

மொழிகளில் இலக்கணத் திருத்தத்திற்கு மூன்று நெறிமுறைகள் பயன்படுத்தப்படுகின்றன; அவையாவன:

(1) தொடரியல் அடிப்படையிலான இலக்கணத்திருத்தம் (syntax-based checking)

(2) புள்ளியியல் அடிப்படையிலான இலக்கணத்திருத்தம் (Statistics-based checking)

(3) விதி அடிப்படையிலான இலக்கணத்திருத்தம் (Rule-based checking)

2.2.1 தொடரியல் அடிப்படையிலான திருத்தம் (Syntax based checking)

தொடரியல் அடிப்படையிலான இலக்கணத் திருத்தத்தில் ஒவ்வொரு வாக்கியமும் அவற்றின் இலக்கணச் செம்மைக்காக முழுமையாகப் பகுத்துக்குறிக்கப்படும் (parsed). தொடரியல் பகுத்துக்குறித்தல் பிழையானால் பனுவல்/உரை தவறானதாகக் கருத்தப்படும். ஜென்சன் (Jensan 1993) என்பவரால் உருவாக்கப்பட்ட அணுகுமுறையயில் ஒரு உரை முழுவதும் பகுத்துக்குறிக்கப்படும். அதாவது வாக்கியங்கள் ஆயப்பட்டு ஒவ்வொரு வாக்கியத்திற்கும் ஒரு கிளை அமைப்பு தரப்படும். பகுத்துக்குறித்தல் வெற்றிகரமாக நடைபெறாவிட்டால் உரையானது தவறானதாகக் கருதப்படும்.

தொடரியியல் அடிப்படையிலான அணுகுமுறையின் சிறப்பு என்னவென்றால் இலக்கணம் முழுமையாக இருந்தால் இலக்கணத் திருத்தலும் முழுமையாக அமையும். அதாவது இலக்கணத் திருத்தி தவறு எவ்வளவு தெளிவற்று இருந்தாலும் எந்தத் தவறான வாக்கியத்தையும் கண்டுபிடிக்கும். துரதிஷ்ட வசமாக

இலக்கணத் திருத்தி தவறான வாக்கியங்களை மட்டும் தெரிந்துகொள்ளும்; அது பயன்படுத்துவோருக்குச் சரியாக என்ன சிக்கல் என்பதைக் கூற இயலாது. இதற்காகக் கூடுதல் விதிகள் தேவைப்படும்; அதுவும் தவறான வாக்கியங்களைப் பகுத்துக்குறிக்கும். ஒரு வாக்கியத்தை அம்மாதிரியான கூடுதல் விதியால் மட்டும் பகுத்துக்குறிக்க இயலுமென்றால் அவ்வாக்கியம் தவறானதாகும். இந்த நுட்பம் கட்டுப்பாடு நெகிழ்வு (Constraint Relaxation) எனப்படும்.

எனினும், தொடரியல் சார்ந்த அணுகுமுறையில் ஒரு முக்கிய பிரச்சனை உள்ளது: அதற்கு ஒருவர் சரிபார்க்க விரும்பும் அனைத்து வகையான பனுவல்களையும் உள்ளடக்கிய ஒரு முழுமையான இலக்கணம் தேவை. பல இலக்கண கோட்பாடுகள் இருந்தாலும் இன்னும் வலுவான பரந்த உள்ளடக்கப்பரப்புப் பாகுபடுத்தி (robust broad-coverage parser) பலரறிய எதுவும் கிடைக்கவில்லை. மேலும் பாகுபடுத்திகள் இயற்கையான பொருண்மை மயக்கத்தால் பாதிக்கப்படுகின்றன; இதன் காரணமாகச் சரியன வாக்கியங்களுக்கும் வழக்கமாக ஒன்றுக்கும் மேற்பட்ட விளைவுகள் திரும்பக்கிடைக்கின்றன.

2.2.2. **புள்ளியியல் அடிப்படையிலான திருத்தம்** (Statistical based checking)

புள்ளியியல் அடிப்படையிலான இலக்கணத்திருத்தத்தில் சொல்வகைப்பாடு அடையாள (POS tag) வரிசைகள் ஒரு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியிலிருந்து உருவாக்கப்படும்; இவ்வரிசைகளின் நிகழ்வெண்

அதாவது நிகழ்தகவு/சாத்தியக்கூறு (probability) குறித்துக்கொள்ளப்படும். சொல்வகை அடையாளப்படுத்தப்பட்ட உரை/பனுவல் சில நிகழ்வெண்ணைக் காட்டிலும் குறைவான நிகழ்வெண்ணில் சொல்வகைப்பாடு வரிசைகளைக் கொண்டிருந்தால் அப்பனுவல் பிழையானதாகக் கருத்தப்படும். புள்ளியியல் அடிப்படை அணுகுமுறை முக்கியமாக அடையாளப்படுத்தப்பட்ட பயிற்சி தரவுத்தொகுதியிலிருந்து (tagged training corpus) விதிகளைக் கற்றுக்கொள்ளும். விதி அடிப்படை அணுகுமுறையில் விதிகள் கையால் எழுத்தப்படவேண்டும் என்பது தவிர அணுகுமுறை புள்ளியியல் அணுகுமுறையுடன் ஒற்றுமை உடையதாகும்.

இருப்பினும் ஆங்கிலத்திற்கு மிகவும் பயன்படுத்தப்படுகின்ற இலக்கணத்திருத்திகளில் ஒன்றான மைகிரோசாஃப்ட் ஆபீஸ் சூட் (Microsoft Office Suit) இலக்கணத் திருத்தியும் எதிர்கருத்திற்கு (controversy) அப்பாற்பட்டதல்ல. தற்போது இலக்கணத் திருத்தி உருவாக்கம் மிகவும் எளிமையான செயல் அல்ல என்பதை நிரூபிக்கின்றது.

இது அட்வல் (Attwell1987) என்பவரால் உருவாக்கப்பட்ட அணுகுமுறையில் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்தப்பட்ட தரவுத் தொகுதி சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட கோர்வைகளின் பட்டியலை உருவாக்கப் பயன்படுத்தப்படும். சில கோர்வைகள் பொதுவானதாகும்.

எ.கா.

அடைகொளி அடை + பெயரடை+ பெயர் = The old man.

பிற எல்லாவற்றிலும் நேர்வதில்லை.

எ.கா.

அடைகொளி அடை + பெயரடை

தரவுத் தொகுதியில் அடிக்கடி நேரும் கோர்வைகளைப் பிற உரைகளில் சரியானதாகக் கருத இயலும்; அரிதாக நேரும் கோர்வைகள் தவறானதாகக் கருதப்படும்.

புள்ளியியல் அணுகுமுறையில் என்கிராம் பகுப்பாய்வை (n-gram analysis) பயன்படுத்தி ஒரு வாக்கியத்தின் சாத்தியக்கூறை அளக்கவியலும். எடுத்துக்காட்டாக, He is playing என்ற வாக்கியத்தின் பைகிராம் சத்தியக்கூறு பின்வருமாறு அமையும்.

P(“He is playing.”) = P(He|<start>)*P(is|he)*P(playing|is)*P(.|playing)

இம்மூன்று சொற்களில் ஏதேனும் ஒன்று பயிற்சித் தரவுத்தொகுதியில் இல்லாவிடில், வாக்கியத்தின் சாத்தியக்கூறு, பெருக்குத்தொகையின் காரணமாகப் பூஜியமாக மாறும். நாம் இப்புள்ளியியல் நெறிமுறையில் சொற்களைக் கருத்திற்கொண்டால் மொழியின் எல்லாச் சொற்களையும் கொண்டுள்ள பெரிய தரவுத்தொகுதி நமக்குத் தேவைப்படும்.

இச்சிக்கலைத் தீர்ப்பதற்காக நாம் தனிச்சொற்களுக்குப் பதிலாகச் சொல்வகைப்பாட்டு அடையாளங்களைப் (POS tags) பயன்படுத்த இயலும். வேறுபாடு என்னவென்றால், முன்னர் நாம் எச்சொற்கள் ஒரு சொல்லுக்குப்பின் வரும் கூடுதல் சாத்தியக்கூறைக் கொண்டிருக்கின்றன என்பதைப் பரிசோதித்தோம்; தற்போது எச்சொல்வகை அடையாளங்கள் ஒரு சொல்வகைப்பாட்டு அடையாளத்திற்குப்பின் வரும் கூடுதல் சாத்தியக்கூறைக் கொண்டிருக்கின்றன என்று பரிசோதிப்போம். நாம் சொல்வகைப்பாடு அடையாளங்களைப் பயன்படுத்தினால், சொற்கள் மாறிகளாகும் (variables).

“He is playing.” என்ற முந்தைய எடுத்துக்காட்டு வாக்கியத்தை எடுத்துக்கொள்ளவும். சொல்வகைப்பாடு அடையாளப்படுத்தலுக்குப் பிறகு அவ்வாக்கியம் He/pps is/bez playing/vbg./.” என மாறும். இப்போது வாக்கியத்தின் சாத்தியக்கூறை கணக்கிட நாம் சொல்வகைப்பாட்டு வரிசையைப் பயன்படுத்த இயலும்.

P(pps bez vbg.) = P(pps|<start>)*P(bez|pps)*P(vbg|bez)*P(.|vbg)

நாம் கூறுகிற இலக்கண திருத்தி பின்வருமாறு செயலாற்றுகின்றது.

1. வாக்கியத்தின் ஒவ்வொரு சொல்லுக்கும் ஒரு அடையாளம் தரவும்.
2. அடையாள வரிசையின் (tag sequence) சாத்தியக்கூறை தீர்மானிக்க என்கிராம் பகுப்பாய்வை பயன்படுத்தவும் (இதில் என்=3 அதாவது டிரை கிராம்)
3. சாத்தியக்கூறு சில நிகழ்வெண்ணிற்கு மேல் இருந்தால் வாக்கியம் இலக்கண அடிப்படையில் சரியானது என்று கருதப்படும். நமது மாதிரியில் சாத்தியக்கூறு பூஜியத்திற்கு அதிகமாக இருந்தால் வாக்கியம் சரியானது என்று கருதப்படும். இரண்டோ அதற்கு மேற்பட்ட அடுத்துவருகிற அடையாளங்கள் பொருந்த இயலவில்லை என்றால் (அதாவது இயைபின்றி இருந்தால்) வரிசையின் சாத்தியக்கூறு பூஜியமாகும்.

முதலில் நமக்கு ஒரு சொல்வகை அடையாளப்படுத்தி (POS tagger) தேவை. அது தானியக்கமாகச் சொற்களை அடையாளப்படுத்தும்; அல்லது நாம் வாக்கியத்தின் சொற்களைக் கையால் அடையாளப்படுத்தவேண்டும். பின்னர் அடையாள

வரிசையின் சாத்தியக்கூறைத் தீர்மானிக்க (இரண்டு முந்தைய அடையாளங்களைப் பார்க்கும்) டிரை கிராம் மாதிரியைப் பயன்படுத்தவும்; இறுதியாக அவ்வடையாள வரிசையின் சாத்தியக்கூறின் அடிப்படையில் இலக்கணச் செம்மையின் தீர்மானத்தைச் செய்யவும்/எடுக்கவும். எடுத்துக்காட்டாக பிரவுண் தரவுத்தொகுதி (Brown Corpus) மற்றும் பிரில்லின் சொல்வகைப்பாடு அடையாளப்படுத்தியை (Brill's tagger), "He saw the book on the table" என்ற வாக்கியத்தின் கணக்கீடுகள் பின்வருவனவாகும்:

He/pps saw/vbd the/at book/nn on/in the/at table/nn ./.
P(pps|None None) = 0.0635486169593
P(vbd|None pps) = 0.213047910296
P(at|pps vbd) = 0.166456494352
P(nn|vbd at) = 0.48308668061
P(in|at nn) = 0.362738953306
P(at|nn in) = 0.350967938144
P(nn| in at = 0.44004695623
P(.|at nn) = 0.0847696646819

அடையாள வரிசையின் சாத்தியக்கூறு = 5.16478478489e-09

இலக்கணத்திருத்தியின் முடிவு = இவ்வாக்கியம் சாத்திய அடிபடையில் சரியானதாகும்.

நாம் உடன்பாட்டு இயைபு இல்லாத "He have the book I want" என்ற வாக்கியத்தை ட்ரைகிராம் அடிப்படையில் ஆய்ந்தால் இலக்கணத் திருத்தி பின்வரும் கணக்கீடுகளைத் தரும்.

He/pps have/hv the/at book/nn I/ppss want/vb./.

P(pps|none none) = 0.0635486169593

P(hv|none pps) = 0.0

P(at|pps hv) = 0.0

P(nn|hv at) = 0.0491712707182

P(ppss|at nn) = 0.00493575681605

P(vb|nn ppss) = 0.293785310734

P(.|ppss vb) = 0.0361445783133

அடையாள வரிசையின் சாத்தியக்கூறு =0.0

இலக்கணத்திருத்தியின் முடிவு = தவறானது அல்லது கண்டுபிடிப்பது சாத்தியமில்லை

புள்ளியியல் சார்ந்த பாகுபடுத்திகள் அவற்றின் முடிவுகளை விளக்குவது கடினம் என்ற இடரைக் கொண்டிருக்கின்றது: ஒரு தவறான எச்சரிக்கைப் பிழை இருந்தால் பயன்பாட்டாளர் குறிப்பிட்ட பிழைச் செய்தி இல்லாத காரணத்தால் ஏன் அவரது உள்ளீடு தவறாகக் கருதப்படுகிறது என்று ஆச்சரியப்படுவர். உருவாக்குபவர் கூட ஒழுங்குமுறையின் தீர்ப்பைப் புரிந்துகொள்ள ஒழுங்குமுறைக்குப் பயிற்சி அளிக்கப்பட்ட தரவுத்தொகுதியை அணுக வேண்டிவரும். மற்றொரு சிக்கல் யாராவது ஒருவர் பொதுவல்லாத மற்றும் தவறான கட்டுகளைப் பொதுவல்லாத ஆனால் சரியான கட்டுகளிலிருந்து பிரிக்கும் தொடக்க நிலையை நிறுவவேண்டும். இந்தச் செயல் 0-வுக்கும் 100-க்கும் இடையில் ஏதாவது மதிப்பை நிறுவும் பயன்பாட்டாளருக்குத் தரப்படவேண்டும். தொடக்க நிலையின் கருத்து (நடை மற்றும் கட்டப்பட்ட இறுதி நேர்வுகள் தவிர) வாக்கியங்கள் வழக்கமாகச் சரி அல்லது தவறு என்ற புலனுணர்வுடன் உண்மையில் இணங்கவேண்டும் என்பதில்லை.

2.2.3. விதி அடிப்படையிலான திருத்தம் (Rule based checking)

இது டேனியன் ராபர் மற்றும் பிறரால் (Denial Naber et al) பயன்படுத்தப்பட்ட அணுகுமுறையில் குறைந்தது சொல்வகைப்பாட்டுக்காக அடையாளப் படுத்தப்பட்ட உரை ஒரு குழும விதிகளாகப் பொருத்தம் பார்க்கப்படும். இந்த அணுகுமுறை புள்ளியியல் அடிப்படையிலான அணுகுமுறைக்கு ஒற்றுமையுள்ளது; ஆனால் உரைகள் மனிதனால் உருவாக்கப்பட்டதாகும்.

பிற அணுகுமுறைகளில் கூறப்பட்ட சிக்கல்கள் காரணமாகக் கண்டிப்பான விதி அடிப்படையிலான ஒழுங்குமுறை இங்கு உருவாக்கப்படும். தொடரியல் அடிப்படையிலான பிழை திருத்திகளைப் போல அல்லாமல் விதி அடிப்படையிலான பிழை திருத்தி ஒரு போதும் முழுமையாக இருப்பதில்லை; அதாவது அதில் கண்டுபிடிக்க இயலாத பிழைகள் இருந்துகொண்டே இருக்கும். இருப்பினும் அதற்கு பல அனுகூலங்கள் உண்டு:

- ஒரு வாக்கியம் சரிபார்க்க முழுமையாக இருக்கத் தேவையில்லை; மாறாக மென்பொருள் பனுவலை அது தட்டச்சு செய்யப்படும் போதே சரிபார்த்து உடனடியான பதில் கருத்தைத் தரவியலும்.

- ஒவ்வொரு விதியும் வெளிப்பாட்டு விளக்கம் கொண்டிருப்பதால் அதற்கு உருவங்கொடுப்பது எளிது மற்றும் தனித்தனியாக முடுக்கவோ அணைக்கவோ இயலும்.

- அது பயனுள்ள கருத்துரைகளுடன் இலக்கண விதிகளை விளக்கியே விளக்கமான பிழைச் செய்திகளைத் தர இயலும்.

- விதி ஒழுங்கமைப்பு குறைந்தது பல எளிய ஆனால் பொதுவான பிழை நேர்வுகளை எளிதில் புரிந்துகொள்ள இயலுமாகையால் அதை அதன் பயன்பாட்டாளர்களுக்கு எளிதாக நீட்சி செய்யவியலும்

- அதைப் படிப்படியாக ஒரு விதியில் தொடங்கி பின்னர் விதிக்கு விதி நீட்சி செய்து உருவாக்கவியலும்.

2.3 நடை மற்றும் இலக்கணத்திருத்தி

நடை மற்றும் இலக்கண திருத்தி (style and grammar checker) ஒரு பனுவலை/உரையை ஏற்கும் பின்னர் சாத்தியமான தவறுகளைத் திருப்பி அனுப்பும். தவறுகளை/பிழைகளைக் கண்டுபிடிப்பதற்காக பனுவலின் ஒவ்வொரு சொல்லுக்கும் அதன் சொல்வகைப்பாடு மதிப்பிடப்படும்/தரப்படும்; ஒவ்வொரு வாக்கியமும் தொடர் உறுப்புகளாக/ தொடர்பன்களாகப் (chunkers) பிரிக்கப்படும். எடுத்துகாட்டாகப் பெயர்த்தொடர்கள் பின்னர் பனுவல் பிழை திருத்தியின் முன்வரையறை செய்யப்பட்ட பிழை விதிகளுடன் ஒப்பிட்டுப்பார்க்கப்படும். ஒரு விதி பொருந்தினால் பொருந்தும் இடத்தில் உரையில் பிழை இருப்பதாகக் ஊகிக்கப்படும்/கருதப்படும். விதிகள் பிழைகளைச் சொற்கள், சொல்வகைப்பாட்டு அடையாளங்கள், தொடர்பன்கள்/தொடர் உறுப்புகள் இவற்றின் ஒருங்கமைப்பாக (pattern) விளக்கும். ஒவ்வொரு விதியும் பிழைக்கு விளக்கத்தை உட்படுத்தும்; அது பயன்படுத்துபவர்க்குக் காட்டப்படும்.

அமைப்பொழுங்கின் விதி அடிப்படையிலான அணுகுமுறை பயன்படுத்துபவர்கள் தங்களது விதிகளை உருவாக்கிக்கொள்ள இயலச்செய்கின்ற

அளவுக்கு எளிமையானதாகும்; பல எடுத்துக்காட்டான பிழைகளைக் கண்டுகொள்வதற்குப் போதுமான அளவு திறன் வாய்ந்ததாகும். விதிகள் எளிய XML வடிவமைப்பில் வெளிப்படுத்தப்படுகின்றது; இது பிழைகளை விளக்குவதுடன் பயன்படக்கூடிய பிழைச் செய்திகளையும் எடுத்துக்காட்டு வாக்கியங்களையும் கொண்டிருக்கின்றது.

2.3.1. கோட்பாட்டுப்பின்னணி

எழுத்துப்பிழை திருத்தம் பயனுள்ளதாகக் கருதப்பட்ட அதே காரணத்திற்காக நடை மற்றும் இலக்கணப்பிழை திருத்தமும் பயனுள்ளதாகும். இது மக்கள் பிழையில்லாமல் ஆவணங்களை எழுதுவதற்குப் பயன்படுகின்றது. நடை மற்றும் இலக்கணப்பிழை திருத்தி பயனுள்ளதாக அமையச் சில தேவைகளைப் பூர்த்திசெய்யவேண்டும்.

- அது விரைவாக இருக்கவேண்டும்; அதாவது இடையீடுப் பயன்பாட்டிற்குப் போதுமான அளவு விரைவாக இருக்கவேண்டும்.
- அது இருக்கின்ற சொல்லாய்வியுடன் நன்றாக இணைக்கப் பட்டிருக்கவேண்டும்.
- சரியான வாக்கியங்களைப் பெரும்பான்மையும் தவறாகக் காட்டக்கூடாது.
- சொந்தத் தேவைகளுக்காக அதை மாற்றிக்கொள்வது சாத்தியப்படவேண்டும்.
- இறுதியாக அது முழுமையாக இருக்கவேண்டும்; அதாவது அது பனுவலில் உள்ள பெரும்பான்மையான பிழைகளைக் கண்டுபிடிக்கவேணடும்.

எழுத்துப்பனுவலில் தோன்றும் பிழைகளைப் பல்வேறு பட்ட வழிகளில் வகைப்படுத்த இயலும்.

எழுத்துப் பிழைகள்: இது எழுத்துப்பிழை திருத்தியால் கண்டுபிடிக்க இயலும் பிழையாகும் என வரையறை விளக்கம் செய்யப்படுகின்றது. இலக்கணப் பிழைதிருத்திகள் அறியப்பட்ட சொற்களின் பெரிய பட்டியலுடன் சொற்களை ஒப்பிட்டுப்பார்க்கும். சொல் அந்தப் பட்டியலில் இல்லாவிட்டால் அது பிழையானது என்று கருதப்படும். ஒற்றுமையுள்ள (சரியான) சொற்கள் பதில் சொற்களாகத் தரப்படும். எடுத்துக்காட்டாக *Gemran என்பது German எனத் திருத்தப்படும்.

இலக்கணப் பிழைகள்: ஒரு மொழியின் இலக்கண விதிகளுடன் ஒத்துப்போகாத ஒரு வாக்கியத்தை விளைவிக்கும் பிழை இலக்கணப் பிழையாகும். எழுத்துப்பிழை திருத்தம் போல அல்லாமல் இலக்கணப்பிழை திருத்தம் சூழல் ததகவலைப் பயன்படுத்த வேண்டியது அவசியமாகும்.

*Harry Potter bigger then than Titanic?

இப்பிழை தட்டச்சால் ஏற்பட்டதா, எழுதுபவரின் மனதில் *then* என்பதற்கும் than என்பதற்கும் இடையில் ஏற்பட்ட தவறான புரிந்துகொள்கையால் ஏற்பட்டதா என்பதைத் தீர்மானிக்க இயலாது. *then* என்பதைப் பயன்படுத்துவது தவறு என்று தெளிவாகத் தெரிவதால் இது இலக்கணப் பிழையாகும்.

இலக்கணப் பிழைகளை அமைப்புசார் பிழை மற்றும் அமைப்புசாரா பிழை என்று பகுக்கலாம். அமைப்பு சார்ந்த பிழைகளைச் சொற்களைச் செருகியோ நீக்கியோ ஒன்றோ இரண்டோ சொற்களை நகர்த்தியோ மட்டுமே திருத்த இயலும். அமைப்புசாராத பிழைகளை இருக்கிற சொல்லை வேற்று சொல்லால் இடம்பெயர்த்து திருத்தம் செய்ய இயலும்.

நடைப்பிழைகள்: இயல்பற்ற/அசாதாராண சொற்களும் சிக்கலான வாக்கிய அமைப்புகளும் உரையை/பனுவலைப் புரிந்துகொள்வதைக் கடினமாக்குகின்றது; இது யாராலும் விரும்பப்படுவதில்லை. இந்நேர்வுகள் நடைப் பிழைகள் என்று கருதப்படும். இலக்கணப்பிழைகளைப் போல் அல்லாமல் நடைப் பிழைகள் அதிகமாகச் சூழலைச் சார்ந்தும் பனுவல்/உரை வகையைச் சார்ந்தும் அமையும். எடுத்துக்காட்டாக, சொந்த நண்பர்களுக்கிடையிலான மின்னஞ்சல் கருதுப்பரிமாற்றம் மொழியின் படைப்பாக்கப் பயன்பாட்டை அனுமதிக்கின்றது; அனால் தொழில் நுட்ப ஆணவப்படுத்தல் பொருண்மை மயக்கத்தால் பாதிக்கலாகாது. அமைப்பு உறுதிப்பாடு (configurability) இலக்கணப் பிழை திருத்தங்களை விட நடை பிழை திருத்தத்திற்குக் கூடுதல் முக்கியத்துவம் வாய்ந்ததாகும்.

எடுத்துக்காட்டாக, பினவரும் வாக்கியத்தை எடுத்துக்கொள்ளவும்.

> But it [a human reason] quickly discover that, in this way, its labours must remain ever incomplete, because new questions never cease to present themselves, and thus it finds itself compelled to have recourse to principles which transcend the region of experiences, while they are regarded by common sense without distrust.

இவ்வாக்கியம் உண்மையான காரணத்தால் காந்தின் விமர்சனத் திறனாய்விலிருந்து (Kant's Critique) விளைந்ததாகும். இது 48 சொல் நீளமுள்ளதாகும்; பலரும் இது புரிந்துகொள்வதற்குக் கடினமானது என்று ஒத்துக்கொள்வர். இதற்குக் காரணம் அதன் நீளம், (transcend போன்ற) கடினமான

சொற்றொகை, இரட்டை எதிர்மறையின் (double negation) பயன்பாடு. இக்காலகட்டத்தில் புரிந்துகொள்வதற்கு எளிமையான ஆணவத்தின் தேவையுள்ள சூழலில், இவ்வாக்கியத்தில் நடைச் சிக்கல் உள்ளதாகக் கருதலாம்.

பொருண்மைப் பிழைகள்: நடைப் பிழையோ இலக்கணப்பிழையோ எழுத்துப் பிழையோ இல்லாத தவறான தகவல் உள்ள ஒரு வாக்கியத்தில் பொருண்மைப் பிழை உள்ளதாகக் கருதலாம். பொருண்மைப் பிழைகளைக் கண்டுபிடிக்க மிகப் பரந்த உலகறிவு (world knowledge) தேவையாதலால், இப்பிழைகளை வழக்கமாகக் கண்டுபிடிக்க இயலாது. எடுத்துக்காட்டாக, “MySQL is a great editor for programming!” என்ற வாக்கியம் உண்மையோ பொய்யோ அல்ல; MySQL என்பது ஒரு எடிட்டர் அல்ல ஒரு டேட்டா பேஸ் ஆகும் என்பதால் இது பொருளைத் தரவில்லை. பரந்த உலகறிவு இல்லாமல் இதைத் தெரிந்துகொள்ள இயலாது. உலகறிவு சூழலின் ஒரு வடிவாகும்; ஆனால் இன்றைய காலகட்டத்தில் இது மென்பொருளின் புரிந்துகொள்ளலுக்கு அப்பாற்பட்டதாகும்.

பிழைகளுக்கும் (errors) தவறுகளுக்கும் (mistakes) வேறுபாடு பாராட்டவேண்டியதில்லை. தவறாகப் கருதப்படுபவைகளும் மோசமாக எழுதப்படுபவைகளும் பிழைகளாகும். இலக்கணம் (அல்லது தொடரியல்) சரியான வாக்கியங்கள் எவ்வாறு இருக்கவேண்டும் என்பதை விளக்கும் விதிகளின் ஒழுங்கமைப்பாகும் (system of rules). எவ்வாறோ விதிகள் மக்கள் மனதில் இருக்கின்றது; இதன் காரணமாகப் பெரும்பாலான மக்கள் ஒரு வாக்கியம் தவறானதா இல்லையா என்பதை எளிதில் தீர்மானிக்க இயலும். உள்ளுணர்வு / மன உணர்வு (intuition) அடிப்படையில் சரியானது/தவறானது என்ற தீர்மானத்தைக் கடினமாக்கும் எல்லையை ஒட்டிவரும் நேர்வுகளைக் காண்பது

சாத்தியமாகும். பிழையான செய்திக்குப் பதிலாக விளக்க எச்சரிக்கைச் செய்தியைக் (descriptive warning message) காட்டி ஒரு மென்பொருள் இந்நேர்வுகளைக் கையாளும். இலக்கண விதி சில சரியான வாக்கியங்களுக்கும் பொருந்துவது தெரிந்தால், எச்சரிக்கைச் செய்திகளும் நடைப் பிழைகளுக்கும் இலக்கணப் பிழைகளுக்கும் பயனுள்ளவையாக அமையும்.

2.3.2 சொல்வகை அடையாளப்படுத்தல் (Parts-of-speech tagging)

சொல்வகை அடையாளப்படுத்தல் (parts-of-speech tagging) ஒவ்வொரு சொல்லுக்கும் அதன் சொல்வகையை அடையாளப் படுத்துவதாகும். எந்தச் சொல்வகை அடையாளங்கள் உள்ளனன என்று அறுதியிட்டு விளக்கப்படவில்லை; அனால் மிகப் பொதுவானவைகள் பெயர், வினை, வரையறை அடை (அடைகொளிஅடை), பெயரடை, வினையடை என்பனவாகும். பெயர்களை ஒருமை, பன்மை என்றும் வினைகளை இறந்தகாலம், நிகழ்காலம், எதிர்காலம் என்றும் இது போன்று மேலும் பகுக்கலாம்.

அதிகமாகச் சொல்வகை அடையாளங்கள் இருந்தால் அது கூடுதல் சிக்கலுக்கு வழிவகுக்கும்; ஏனென்றால் பல சொற்களுக்கு சூழல் அடிப்படையில் வேறுபட்ட சொல்வகைபாடுகள் இருக்கக்கூடும். ஆங்கிலத்தில் பல சொற்களைப் பெயராகவும் வினையாகவும் பயன்படுத்த இயலும்; எடுத்துக்காட்டாக house என்பது ‘கட்டிடம்’ என்றும் ‘பாதுகாப்பு அளி’ என்றும் பொருள்படும். அனைத்துச் சொற்களிலும் 11.5% தான் சொல்வகைப்பாட்டைப் பொறுத்தவரையில் மயக்கம் உடையன; இவை அடிக்கடி நேரும் சொற்கள் ஆகையால் ஒரு பனுவலில் 40%

சொற்கள் மயக்கம் உடையன. சொல்வகைப்பாடு அடையாளப்படுத்துதல் ஒரு மாதிரி சொல்மயக்க நீக்கச் சிக்கலாகும்.

சூழலைக் கருத்தில் கொள்ளாமல் தரப்பட்ட ஒரு சொல்லுக்கு அடிக்கடி நேரும் சொல்வகைப்பாட்டு அடையாளத்தைத் தேர்ந்தெடுத்து ஒருவர் 91% சொற்களுக்கு அவற்றின் சரியான அடையாளத்தைத் தரவியலும். பெரிய தரவுத் தொகுதியின் புள்ளியியல் ஆய்வின் அடிப்படையிலான பெரும்பான்மையான அடையாளப்படுத்திகள் 95-97% துல்லியத்தைக் காட்டுகின்றன (Brill, 1992). கட்டுப்பாட்டு இலக்கணம் (constraint grammar) 99% துல்லியத்திற்குக் கூடுதல் உரிமைகோருகின்றன (Karlsson, 1995). இவ்விழுக்காடுகளைப் புரிந்துகொள்ளுவதில் நாம் விளிப்புணர்வுடன் இருக்கவேண்டும். சில அடையாளப்படுத்திகள் சில மயக்கம் தரும் நேர்வுகளை விட்டுவிட்டு ஒரு தரப்பட்ட சூழலில் ஒரு சொல்லுக்கு ஒன்றிற்கும் மேற்பட்ட சொல்வகைப்பாட்டு அடையாளத்தை திருப்பி அனுப்பும். திருப்பி அனுப்பப்பட்ட சொல்வகை அடையாளங்களில் குறைந்தது ஒன்று சரியானது என்ற சாத்தியங்கள் ஒரு சொல்வகை அடையாளம் திருப்பி அனுப்பப்படுவதை விடக் கூடுதலாகும். வேறுவிதமாகக் கூறினால் ஒருவர் மீட்டழைப்புக்கும் (recall) துல்லியத்திற்கும் (precision) இடையில் வேறுபாடு கருதவேண்டும்.

2.3.3 தொடர் பகுத்தல் (Phrase chunking)

தொடர் பகுத்தல் சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கும் ஒரு முழு இலக்கணப் பகுப்பாய்வுவுக்கும் இடையே அமைந்துள்ளது: சொல்வகைப்படு அடையாளப்படுத்தல் சொல்நிலையில் மட்டுமே செயல்பாடு செய்யும் போது மற்றும் இலக்கணப் பகுப்பாய்வு (பகுத்துக்குறித்தல்) வாக்கியத்திற்கு

கிளையமைப்பு கட்டமைக்கும் போது, தொடர் பகுத்தல் ஒரு வாக்கியத்திலுள்ள சொல் வரிசைகளுக்கு ஒரு அடையாளத்தை ஒதுக்குகிறது. பெயர்த்தொடர் (பெதொ) மற்றும் வினைத்தொடர் (விதொ) என்பன எடுத்துகாட்டான பகுப்புகளாகும். பெயர்த்தொடர் வரையறை அடைகள், பெயரடைகள், பெயர்கள், மாற்றுப்பெயர்கள் ஆகியனவற்றைக் கொண்டிருக்கும். வினைத்தொடர் ஒரு தனிநிலை வினை அல்லது துணை வினை மற்றும் வினையெச்சம் ஆகியவற்றைக் கொண்டிருக்க இயலும். எடுத்துக்காட்டாக the dog, the big dog, the big brown dog என்பன பெயர்த்தொடர்களுக்கு எடுத்துக்காட்டுகளாகும். பெயர்த்தொடர்களின் பட்டியல் எல்லையற்று நீளுமாகையால் கோட்பாடு அடிப்படையில் பெயர்த்தொடர் எல்லையற்று நீளும். இருப்பினும் இங்கு பெயர்த்தொடர் என்று அழைக்கப்படுவது ஒரு எடுத்துக்காட்டு தான்; சொல்வகை அடையாளப்படுத்தல் போல ஒவ்வொருவரும் அவருடைய பகுப்புப் பெயர்களையும் அவற்றின் பொருண்மைகளையும் உருவாக்க இயலும். ஒரு பகுப்பியால் கண்டுகொள்ளப்படும் பகுப்புகள் பெயர்த்தொடரையும் வினைத்தொடரையும் மட்டும் வைத்துக்கொண்டு முழுப் பனுவலையும் உட்படுத்தத் தேவையில்லை; இது சாத்தியமும் இல்லை.

சொல்வகைப்பாடு அடையாளப்படுத்தல் சொல்லில் செயல்படுவது போன்று தொடர் பகுத்தல் சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட பனுவலில் செயல்படும்: எந்தச் சொல்வகைப்பாடு அடையாள வரிசை எந்தத் தொடர் பகுப்புகளைக் கட்டமைக்கும் என்ற கையால் எழுதிய விதிகள் உள்ளன அல்லது நிகழ்வு சாத்தியத் தொடர் பகுப்பி சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட மற்றும் தொடர் பகுக்கப்பட்ட பனுவல்களில் பயிற்சி அளிக்கப்படும். தொடர் பகுத்தல் ஒரு சொல்வகைப்பாடு அடையாளப்படுத்தப்பட

பனுவலை வேண்டுவதால், அதன் துல்லியம் பயன்படுத்தப்பட்ட சொல்வகை அடையாளப்படுத்தியை விட சிறப்பாக இராது. தொடர்பகுத்தலுக்குப் பட்டியலிடப்பட்டுள்ள சிறப்பான தொடர் பகுப்பி ஒரு துல்லியத்தை அடைகின்றது மற்றும் 94% திருப்பி அழைத்தலை அடைகின்றது; இது ஒரு சராசரி சொல்வகை அடையாளப்படுத்தி எய்துவதைவிடக் குறைவாகும்.

2.3.4 இலக்கணப் பிழைகள்

எளிய மொழியான ஆங்கிலத்திற்குக் கூட இலக்கண விதிகள் விரிவானவை. இங்கு விளக்கப்பட்டுள்ள இலக்கண விதிகள் இவ்விதிகளை மீறும் தரவுத்தொகுதியிலிருந்து எடுக்கப்பட்ட வாக்கியங்கள் அடிப்படையிலானது.

2.3.4.1 எழுவாய்-வினை உடன்பாடு

ஆங்கிலத்தில் எழுவாயும் வினையும் எண் மற்றும் இடம் அடிப்படையில் உடன்பாடு காட்டும். எடுத்துகாட்டாக *They is my favourite Candian authors, என்ற வாக்கியத்தில் எழுவாயும் வினையும் எண்ணில் வேறுபடுகின்றன (they = பலர்பால், is = ஒருமை. * He am running for president, என்பதில் எழுவாயும் வினையும் இடம் அடிப்படையில் உடன்படவில்லை (he = படர்க்கை இடம், am = to be என்பதன் தன்மையிடம்)

இது எளிய நேர்வாகும். சொல்வகைப்பாட்டு அடையாளங்களைக் கொண்ட டோக்கன்களின் வரிசைகளாக பனுவலை எடுத்துக்கொள்ளும் விதி அடிப்படையிலான திருத்தியின் பார்வையை எடுத்துக்கொண்டால் பல சிறப்பான நேர்வுகள் உள்ளன:

1. எழுவாயும் வினையும் பிரிக்கப்பட்டுள்ளன, அதாவது வினை எழுவாய்க்கு நேரடியாகப் பின்னர் வருவதில்லை: *The characters in Shakespear's Twelth Night lives in a world that has been turned upside-down.

2. எழுவாய் ஒரு கூட்டுநிலை எழுவாயாக இருக்கவியலும்: *Chrisie and Prin is characters form Laurence's The diviners.

3. நூலின் தலைப்புகள் எல்லாம் ஒருமை: *Salman Rushdie's Midnight's children are my favourite novel.

2.3.4.2 குறிப்பில்லா அடை மற்றும் தொடர்ந்துவரும் சொல்

குறிப்பில்லா அடையைத் தொடர்ந்துவரும் சொல்லின் உச்சரிப்பு உயிர் ஒலிப்பு முதலில் வரும் படி இருந்தால் a அடைக்குப் பதிலாக an என்ற அடை பயன்படுத்தப்படவேண்டும்.

எ.கா.

A test, a car, a long talk

An idea, an uninteresting speech, an earth quake

2.3.4.3 தொடுவினா

பேச்சு மொழியில் தொடுவினா (tag question) ஒரு கூற்றுவாக்கியத்திற்கு உறுதியைப் பெறுவதற்குப் பயன்படுத்தப்படும். இது ஆங்கிலத்தில் துணைவினையின் எதிர்மறை வடிவையும் வாக்கியத்தின் இறுதியுடன் வாக்கியத்தின் எழுவாயைச் சேர்ப்பதாலும் கட்டமைக்கப்படுகின்றது. எடுத்துக்காட்டாக, *It's warm today* என்பது *it is warm today, isn't it?* என வரும். தொடுவினாக்கள் மின் அஞ்சலில் பயன்படுத்தப்படுகின்றது. ஜெர்மானி தாய்மொழி பேசுபவர்களுக்கு ஆங்கித்திலுள்ள தொடுவினா அமைப்பை புரிந்து கொளவது கடினம். எனவே அவர்கள் ஆங்கிலத் தொடுவினா வடிவை ஜெர்மானிய மொழிக் கட்டமைப்பின் படி தவறாகப் பயன்படுத்துவர்.

2.3.4.4 பிற பிழைகள்

பெரும்பாலான பிற பிழைகள் தொழிநுட்ப அடிப்படையில் இலக்கணப் பிழைகள்; அவைகள் தட்டச்சால் உருவாக்கப்படுகின்றது. பெரும்பாலும் பிழைகள் அது பனுவலைத் திருத்தும்போது சில சொற்களை விட்டுவிடுவதால் ஏற்படுகின்றது.

எ.கா.

*Someone suggested said that it worked for him after he updated to Kernel 2.4.20.

இவ்வாக்கியத்திற்கு உரியவர் *said* என்பதை *suggested* என்பதால் இடம்பெயர்க்க விரும்புகிறார்; ஆனால் *said* என்பதை நீக்குவதற்கு மறந்துவிடுகிறார்.

பெரும்பாலும் ஒன்றுபோல் இருக்கின்ற சொற்கள் கலக்கப்படுகின்றன; மற்றும் எழுத்தாளர் தட்டச்சர் பிழை செய்தாரா அல்லது அவருக்கு வேறுபாடு தெரியவில்லையா என்று கூற இயலாது.

2.3.4.5 வாக்கிய எல்லை கண்டுபிடிப்பு

இலக்கணத்தன்மை வாக்கியங்களைக் குறிப்பிடும். பின்வரும் இரண்டு வாக்கியங்களின் இலக்கணத்தில் பிழை உள்ளது; *Peter* என்ற பெயருக்கும் *she* என்ற பதிலீட்டுப்பெயருக்கும் உடன்பாடு இல்லாததால் அவைகளுக்கிடையில் இலக்கணப் பிழை உள்ளது.

Peter leaves his house. She feels good.

இம்மாதிரியான பிழைகளைக் கையாண்டு திருத்துவது கடினம். இது பொருண்மை நிலையில் வரும் பிழையாகும். மனிதர்களுக்கு இருக்கும் இம்மாதிரியான உல அறிவைக் கணிப்பொறிக்குப் பகிர்ந்து தருவது கடினம்.

சுருக்கச் சொற்கள், எண்கள், அயற்கூற்று என்பன கணிப்பொறிகளுக்குக் கடினமானது. வாக்கர் மற்றும் பிறர் (Walker et al, 2001) வாக்கிய எல்லை

கண்டுபிடிப்புக்கு மூன்று அணுகுமுறைகளை மதிப்பீடு செய்கின்றார். அவர்களின் கவனக்குவிப்பு தங்கள் ஒழுங்கமைப்பில் வாக்கிய எல்லைக் கண்டுபிடிப்பு தேவைப்படுகின்ற தானியங்கு மொழிபெயர்ப்பு ஒழுங்குமுறைகளாகும். இம்மூன்று அணுகுமுறைகள் பின்வருவனவாகும்:

- சொற்களையும் நிறுத்தற்குறிகளையும் தவிர உயர்நிலை விளக்கங்களைப் பயன்படுத்தாமல் மொழிபெயர்ப்பு ஒழுங்குமுறையில் நேரடியாக நடைமுறைப் படுத்தல்: இந்த ஒழுங்குமுறை சுருக்க அகராதியைப் (abrivation lexicon) பயன்படுத்துகின்றது. இதை நடைமுறைப்படுத்துவது சீரான வெளிப்பாடுகளால் (regular expressions) ஊக்கப்படுத்தப்பட்டுள்ளது; ஆனால் எல்லாம் தரப்பட்டுள்ள ஒரு வழியமைப்பு மொழியில் நேரடியாக நடைமுறைப்படுத்தப்பட்டுள்ளது.

- சீரான வெளிப்பாடுகளாக வாக்கியங்களின் விதி அடிப்படையிலான உருப்படுத்தம். சீரான வெளிப்பாடுகள் கூற்றுவழியில் தேவையான அறிவை குறியாக்கம் செய்ய அனுமதிக்கின்றன. இது வெளிப்படையாகக் கூறப்படாவிட்டாலும் இந்நெறி ஒரு சுருக்கச்சொல் அகராதியையும் பயன்படுத்துவதாகத் தெரிகின்றது.

- அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியில் பயிற்சியளிக்கப்பட்ட இயந்திரம் கற்கும் வழிமுறை வரைவின் பயன்பாடு. இவ்வழிமுறை வரைவு தலைப்பெழுத்தாக்கம் மற்றும் சுருக்கச்சொல் அகரதி போன்ற சாத்தியமான வாக்கிய எல்லையின் பல பண்புக்கூறுகளை அளவிடுகின்றது.

வாக்கர் மற்றும் பிறரின் மதிப்பீடு இயந்திரம் கற்றல் வழிமுறைவரைவு சுமார் 98% துல்லியத்தையும் திரும்பப் பெறுதலையம் தருகின்றது என்று காட்டுகின்றது.

எனவே மிக நல்ல எல்லை கண்டுபிடித்தலுடன் கூடிய ஒரு நடை மற்றும் இலக்கண பிழைதிருத்தி வாக்கிய எல்லைகளில் கிட்டத்தட்ட 2% பிழை விகிதத்துடன் செயல்படும். இது தவறாகப் தொடர்பகுதி பகுக்கப்பட்ட வாக்கியங்களின் மேல் செயல்புரியும்படி எதிர்பார்க்கப்படுகின்ற பகுத்துக்குறிப்பானின் மோசமான சிக்கலாகும்; ஏனென்றால் தவறாகச் சேர்க்கப்பட்ட வாக்கிய எல்லையும் தவறப்பட்ட வாக்கிய எல்லையும் பகுத்துக்குறிக்க இயலாத உள்ளீட்டிற்குக் கொண்டு செல்லும். விதி அடிப்படையிலான ஒழுங்குமுறைக்கு இது குறைந்த சிக்கல் உள்ளது: வாக்கிய எல்லைகளை வெளிப்படையாகக் குறிப்பிடும் விதிகளே பாதிக்கப்படும். இவ்விதிகள் வாக்கிய எல்லைகள் தவறாகச் சேர்க்கப்படும் போது தூண்டப்படும் மற்றும் வாக்கியக் கண்டுபிடிப்பால் வாக்கிய எல்லை தவறப்படும் போது அவை தூண்டப்படாது இருக்கும்.

2.3.5 கட்டுப்படுத்தப்பட்ட மொழித் திருத்தம்

கட்டுப்படுத்தப்பட்ட மொழி என்பது மொழியை எளிமையானதாகவும் புரிவதற்கு எளிதாகவும் ஆக்குவதை நோக்கமாகக் கொண்ட விதிகளினால் கட்டுப்படுத்தப்பட்ட இயற்கை மொழியாகும். மற்றொரு வழியில் கூறினால் கட்டுப்படுத்தப்பட்ட மொழி இயற்கைமொழியின் ஒரு துணைநிலை மொழியாகும். கட்டுப்படுத்தப்பட்ட மொழியை உருவாக்குவதன் முக்கியமான நிலைப்பாடு எல்லா மொழியியல் நிலைகளிலும் மயக்கத்தை விலக்குவதாகும். மயக்கமில்லாத மொழிகளுக்கு பொதுவான

இயற்கைமொழிகளோடு ஒப்பிடுகையில் இரண்டு முக்கியமான அனுகூலங்கள் உள்ளன.

- குறிப்பாக இயற்கைமொழியின் இயல்பான பேசுபவர்களாக இல்லதிருக்கையில், அது புரிவதற்கு எளிதாகும். இயல்பான மொழி பேசுபவர்களுக்கும் தவறாகப் புரிதல் மிகவும் குறைக்கப்படும். இது ஆவணங்களைப் படிப்பதை எளிதாக்குவதுடன் சில வகை ஆவணங்களில் மிக முக்கியமானதாகும்; எடுத்துக்காட்டாக வானவூர்தி இயந்திரங்களைப் பாதுகாப்பது பற்றி விளக்கும் ஆவணங்கள்.

- இது கணிப்பொறியால் பகுத்துக்குறிப்பதற்கு எளிதாகும்; எனவே தானியக்கமாக மொழிபெயர்ப்பதற்கு எளிதாகும். இது மொழிபெயர்ப்புச் செயன்முறையில் மிக சிக்கனமானதாகும்; இது மனித முயற்சியால் செய்யப்படும் போது அதிக செலவீனமுள்ளதாகும்.

கட்டுப்படுத்தப்பட்ட மொழியின் எடுத்துக்காட்டான பயன்பாடுகள் பின்வருமாறு அமையும்:

- **சொல் கட்டுப்பாடுகள்**: எடுத்துக்காட்டாக *try* என்ற சொல்லை வினையாக மட்டும் பயன்படுத்தவும்; பெயராகப் பயன்படுத்தவேண்டாம். இது பொருண்மையியல் கட்டுப்பாட்டை குறிப்பாக உணர்த்தும்; ஏனென்றால் அது தொடரியல் கட்டுப்பாடாக வெளிப்படுத்தப்பட்டால் அது *try* என்ற பெயரை நோரிடும் போது ஒரு பிழைச் செய்தியை தூண்டும் ஒரு விதியுடன் காணப்படும்.

- **இலக்கணக் கட்டுப்பாடுகள்**: எடுத்துக்காட்டாக வினையை ஏவல் வடிவில் எழுதவும்.

  Remove oil and grease with a degreasing agent.

  Oil and grease are to be removed with a degreasing agent.

- **பொருண்மை கட்டுப்பாடுகள்**: எடுத்துக்காட்டாக noise என்ற சொல்லை "unwanted sound" என்ற பொருளில் மட்டும் பயன்படுத்தவும்.

- **நடை கட்டுப்பாடுகள்**: ஒரு விதி வாக்கியங்களை 20 சொற்களுக்குள் வைத்துக்கொள்ளவும் ஒரு வாக்கியத்திற்கு ஒரு தலைப்பு எனக் கொள்ளவும் கோரலாம்.

2.3.6 நடை திருத்தம்

இயற்கைமொழிகளின் இலக்கணம் சரியான வாக்கியங்களுக்கும் தவறான வாக்கியங்களுக்கும் இடையிலான வேறுபாட்டைத் தெளிவாகத் தருகின்றது. இது மொழி நடைக்குப் பொருந்தாது. ஒவ்வொ எழுத்தாளரும் எந்த நடை எந்த இடத்தில் விருப்பத்தேர்வு செய்யப்படவேண்டும் என்று தமது சொந்தத் தீர்மானங்களை எடுக்கின்றனர். பலரால் ஒரு ஆவணம் இணைந்து உருவாக்கப்படும் போது இது இணக்கம்/இயைபு இல்லாத நடைக்குக் கொண்டுசெல்லும். இம்மாதிரியான சூழல்களில் நடை திருத்தம் மிகப் பயனுள்ளதாகும். நடையைப் பற்றிய கட்டுப்பாடுகளும் விதிகளும் விரிந்த எல்லையிற் படும் விசயங்களை உட்படுத்தும்.

- **சொற்களின் விருப்பத்தேர்வு**: எல்லோரும் புரிந்துகொள்வதற்குக் கடினமாகையால் வேற்றுமொழியின் சொற்கள் நிராகரிக்கப்படலாம். நடை

திருத்தி அச்சொல்லை அதே பொருள்கொண்ட இயல் மொழிச் சொல்லால் இடம்பெயர்க்கும் படி பரிந்துரைக்கலாம்.

- **எளிமை**: புரிந்துகொள்ளும் எளிய வாக்கிய அமைப்புகளுக்குக் கொண்டு செல்லும் படி வாக்கியங்களின் நீளம் குறைக்கப்படலாம்.

- **நிறுத்தற்குறிகள்**: நிறுத்தற்குறிகளுக்குப் பின்னர் இடம் விடப்படலாம்; ஆனால் நிறுத்தற்குறிகளுக்கு முன்னர் இடம் விடப்படுவதில்லை.

- நாட்களும் நேரங்களும் எண்களும் ஒரு குறிப்பிட்ட நடை வடிவிலேயே எழுத்தப்படவேண்டும். எடுத்துக்காட்டாக 5/7/1999 என்று எழுதுவது மயக்கம் தரும்; மாறாக அது 1999-07-05 என எழுதப்படவேண்டும்.

- *Don't* போன்ற சுருக்கவடிவங்கள் நிராகரிக்கப்பட்டு *do not* என்பதால் இடம்பெயர்க்கப்படலாம்.

- ஒரு குறிப்பிட்ட பெயரை அடிக்கடி பயன்படுத்தும் பனுவல் வழக்கமற்றதாகத் தோன்றலாம். பதிலாக பதிலீடு பெயர்கள் அல்லது ஒருபொருள் பன்மொழிகள் அந்தப் பெயரை குறிப்பிட பயன்படுத்தப்படவேண்டும். தொழில்நுட்ப ஆணவங்களில் தெளிவுக்கும் மயக்கமின்மைகளுக்கும் முக்கியத்துவம் தரப்படவேண்டும்.

2.4. முடிவுரை

எழுத்து மற்றும் இலக்கணத் திருத்தத்திற்கான சாத்தியங்களையும் வழிமுறைகளையும் கோட்பாடுகளையும் கொள்கைகளையும் பற்றி இவ்வியலில்

விளக்கப்பட்டுள்ளது. இதன் அடிப்படையில் தமிழுக்கான எழுத்து மற்றும் இலக்கணத் திருத்தி உருவாக்குவதற்கான சாத்தியங்கள் ஆயப்படும்.

இயல் 2

இயந்திர மொழிப்பெயர்ப்பு

2.1 அறிமுகம்

இன்றைய அறிவியல் உலகத்தின் நாயகன் கணிப்பொறி. இருபதாம் நூற்றாண்டில் அறிமுகப்படுத்தப் பட்ட பல்வேறு பயன்பாட்டுச் சாதனங்களைப் பின் தள்ளிவிட்டு நமக்கு இன்றியமையாத பங்காளியாய் மாறியுள்ளது கணிப்பொறி. வெறும் கணிக்கும் பொருளாக இருந்த கணிப்பொறி தரவுகள் சேமித்து வைக்கும் கருவியாகவும் தகவல் தொடர்பு கருவியாகவும் தரவுளைத் தானே ஆய்ந்து நமக்கு முடிவுகளை வெளிப்படுத்தும் கருவியாகவும் பயன்படுகிறது. மொழிச் செயல்பாட்டிற்கும் கணிப்பொறியின் பங்கு நம்மை வியப்படையச் செய்யும் விதத்தில் வளர்ந்துள்ளது.

தற்காலக் கணிப்பொறியின் நினைவகம் மிகப்பெரிய அளவில்ல் உள்ளது. துணை நினைவகங்களும் மிகப்பெரிய சேகரிப்பு இடத்தைக் கொண்டுள்ளன. பல கோடிக்கணக்கானச் சொற்களால் அமைந்த தரவுகளைக் கணிப்பொறியில் (நினைவகம், பதிவுத்தட்டு, பதிவுநாடா போன்றவற்றில்) சேகரித்து வைத்துத் தேவையாபோது பயன்படுத்திக் கொள்ள இயலும்.

மொழிக்கான தரவுகள் பல வகைகளில் அமையலாம். சொற்களின் தொகுப்பாகவோ, வாக்கியங்களின் தொகுப்பாகவோ அமையலாம். இவைகள் உள்ளடக்கிய ஒரு உரையாகவும் (text) அமையலாம். வெவ்வேறு வகைப்பட்ட உரைகளிலிருந்து தேர்ந்தெடுக்கப்பட்ட பகுதிகளின் தொகுப்பாக அமையலாம். இத்தகைய தரவு விரிதரவவு (Corpus) என்று அழைக்கப்படும். தமிழுக்கான விரிதரவைச் சேகரித்து வைப்பது மிக இன்றியமையாதப் பணியாகும். பல

இலக்கியக் கூறுகளை உள்ளடக்கிய உரைகள் வட்டார வழக்கு, கிளைமொழி வழக்கு, சமுதாய வழக்கு, பேச்சு வழக்கு, இலக்கிய வழக்கு போன்ற பல்வேறு மொழி வழக்குகளை உள்ளடக்கிய உரைகள் எல்லாம் ஒருசேர விரிதரவாக அமையும் இத்தகைய விரிதரவை இருவகையாகப் பிரிக்கலாம்.

1. எழுத்து வடிவில் அமைந்த தரவுத்தொகுதி (Written corpus)
2. பேச்சு வடிவில் அமைந்த தரவுத்தொகுதி (Speech corpus)

இத்தகைய தரவுத்தொகுதிகள் மொழியாய்வுக்கும் மொழிக்கான பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பேருதவி புரிகின்றன.

தரவுகள் கணிப்பொறியால் ஆயப்பட்டுத் தரப்படும். மொழித் தொடர்பான தகவல்களைப் பல வகைப்பட்ட நினைவகத்தில் (வனதட்டு, குறுந்தட்டு, பதிவு நாடா போன்றவற்றில்) சேமித்து வைத்து தேவைப்படும் போது பயன்படுத்திக் கொள்ளலாம். பல்வேறு விதமான தகவல்ககள் இணையத்தளத்தில் நிரம்பிக் கிடக்கின்றன். தமிழ்மொழியின் கட்டமைப்பைக் கணிபபொறி வழி மொழியின் எல்லா நிலைகளிலும் (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) ஆய்ந்து வைத்திருந்தால் தமிழ் மொழியைப் பயன்படுத்தி தகவல்களைப் பெற உதவும். தகவல் தேடும் பொறி (Search engine) தகவல்களைப் பெற உதவும் முக்கிய கருவியாகும். தமிழுக்கு இத்தைகைய தகவல் தேடும் பொறி பெருமளவில் வெற்றியும் அடைந்துள்ளது.

மொழி உரைகளைப் படிநிலை (ஒலியனியல் நிலை, உருபனியல் நிலை, தொடரியல் நிலை, பொருண்மையியல் நிலை) அமைப்பில் வரும் மொழி அலகுகளாகப் பிரித்தாய்ந்து மொழியின் கட்டமைப்பைப் புரிந்து கொளள கணிப்பொறயைப் பயன்படுத்தலாம். பின்வரும் தமிழ் வாக்கியத்தை எடுத்துக் கொள்வோம்.

எ.கா.

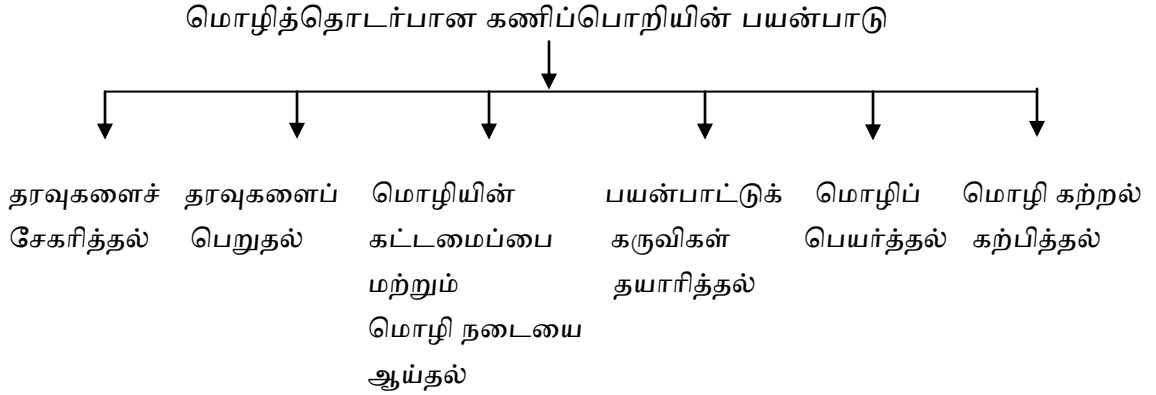
பள்ளி மாணவன் ரகுமான் மரத்திலிருந்து பெரிய பலாப்பழத்தைக் கவனமாகப் பறித்தான்.

பள்ளி மாணவன் ரகுமான் என்பது பெயர்த்தொடர் என்றும், *பலாப்பழத்தைக் கவனமாகப் பறித்தான்* என்பது வினைத்தொடர் என்றும், *பள்ளி, மாணவன், மரம், பலாப்பழம், ரகுமான்* என்பன பெயர்கள் என்றும், பறி என்பது வினை என்றும், *பெரிய* என்பது பெயரடை என்றும், கவனமாக என்பது வினையடை என்றும், இதிலிருந்து என்பது பின்னுருபு என்றும், ஐ-என்பது வேற்றுமையுருபு என்றும் *–த்-* என்பது இறந்தகால இடைநிலை என்றும் *-ஆன்* என்பது படர்க்கை ஆண்பால் வினைமுற்று விகுதி என்றும் பகுப்பாய்வு செய்யலாம். மொழியின் கட்டமைப்பைப் பற்றிய தகவல்களைச் செயற்கை அறிவாகக் கணிப்பொறிக்கு ஊட்டினால் கணிப்பொறி மேற்கண்டது போன்ற வாக்கியங்களைப் பகுப்பாய்வுச் செய்து தந்து விடும்.

ஒரு கால கட்டத்தில் மொழியலார் கள ஆய்வு செய்து தகவலாளிகளின் உதவயால் ஆயப்படாத மொழிகளின் தரவுகளைச் சேகரித்து அதன் மீத சில மொழியியல் கொள்கைகளையும் கோட்பாடுகளையும் பயன்படுத்தி சில கருதுகோள்களின் அடிப்படையிலும் உற்றுநோக்கல் மூலமும் அம்மொழிகளின் ஒலியன், உருபன், தொடர், பொருண்மை அமைப்புகளைக் கண்டுபிடிக்க முயலுவர். இது பண்புசார் மொழியியல் (Qualitative linguistics) எனப்படும். இப்போது கணிப்பொறியின் வரவால் அத்தகைய நிலை மாறி மொழிகளின் தரவுத்தொகுதிகள் உருவாக்கப்பட்டு அதன் மீது புள்ளியியல் நெறிமுறைகளின் அடிப்படையில் மேற்சொன்ன மொழி அமைப்புகள் கண்டுப்பிடிக்கப்படும். இதை அளவுசார் மொழியியல் (Quantitive Linguistics) என்பர். தற்போது மொழி ஆய்வுகள் எல்லாம் கணினியைப் பயன்படுத்தி தரவுத்தொகுதிகள் மீது செய்யப்படும் ஆய்வுகளாக மாறியள்ளன.

மொழி உரையை இடுவிரலாகத் (input) தந்து தக்க வழியமைப்புகளும் (Programs)கட்டளைகளும் தந்தால் மொழியின் உரைகளை வாக்கியங்களாகவும் வாக்கியங்களை தொடர்களாகவும், தொடர்களைச் சொற்களாகவும் சொற்களை ஒலியன்களாகவும் ஒலியன்களை ஒலிக் கூறுகளாகவும் பிரித்து ஆயலாம். மேலும் மொழி அலகுகளக்கு இடையே உள்ள பொருண்மை உறவுகளையும் ஆயலாம். இதற்கானத் தொடரியல் ஆய்வி (Syntactic analyzer), உருபனியல் ஆய்வி (Morphological analyzer), ஒலியனியல் ஆய்வி (Phonological analyzer), பொருண்மையியல் ஆய்வி (Sematic analyzer) என்ற ஆய்விகளைத் தயாரித்து பயன்பாட்டில் பயன்படுத்துகின்றன.

மொழித்தொடர்பானக் கணிப்பொறியின் பயன்பாட்டை ஆறு முக்கிய அம்சங்களாகப் பகுக்கலாம்.



இவ்வமைப்பின் படி மொழி ஆய்வில் கணிப்பொறியின் பங்கு வியக்கும் அளவில் இருக்கிறது என்பதை நம்மால் உணரமுடிகிறது. இக்காலத்தில் எந்த செயலைச் செய்வதற்கும் கணினி இல்லாமல் செய்ய முடியாது. அந்த அளவிற்கு கணினியின் பயன்பாடு பேருதவி புரிந்து வருகின்றது. மொழியாயவில் கணிப்பொறி முதல் கருவியாகக் கொண்டு விரைவில் ஒரு ஆய்வை குறித்த நேரத்தில், குறிப்பிட்ட காலத்தில் செய்து முடிப்பதற்கு கணினியின் பங்கு முக்கியமானதாகக்

கருதப்படுகிறது.

இயந்திர மொழி பெயர்ப்பு இன்றைய கணினி யுகத்தில் மிக முக்கியமானதாகக் கருதப்படுகிறது. பல நூற்றாண்டுகளாக மனிதனால் செய்யப்பட்ட மொழி பெயர்ப்பு செயல்பாட்டைக் கணினியால் எவ்வாறு எளிதாகவும் விரைவாகவும் செய்யலாம் என்ற முயற்சிகள் 1950-களிலேயே தொங்கப்பட்டு பின்னர் கைவிடப்பட்டு மீண்டும் புதிய திசையில் வளர்ந்து கொண்டிருக்கிறது. கணிப்பொறியின் தொழில் நுட்பத்தில் ஏற்பட்ட முன்னேற்றங்கள் இயந்திர மொழி பெயர்ப்பிலும் முன்னேற்றத்தை ஏற்படுத்தியது, பலவிதமான மாதிரிகள் அடிப்படையில் இயந்திர மொழி பெயர்ப்பு செய்யப்பட்டு பரிசோதனைக்கு உள்ளாக்கப்பட்டது. நேரடியான சொல்லிலிருந்து சொல் மொழிபெயர்ப்பு மாற்றம், தொடரிலிருந்து தொடர் மொழி பெயர்ப்பு மாற்றம், பொருண்மையிலிருந்து பொருள் மொழி பெயர்ப்பு மாற்றம் எனப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

கணிப்பொறியை மட்டும் வைத்துக்கொண்டு மொழிப்பெயர்ப்பு செய்ய இயலாது என்று புரிந்துகொள்ளப்பட்டதன் விளைவாகக் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு, மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு, முற்றிலும் தானியங்கி மொழி பெயர்ப்பு எனப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

பலவிதமான இலக்கண வடிவமைப்புகள் (Grammatical Formalisms) மொழி பெயர்ப்பிற்குப் பயன்படுத்தப்பட்டன. எடுத்துக்காட்டாக முற்றுநிலைத் தானியங்கி, சூழல் கட்டிலா இலக்கணம் (Context Freee Grammar) சூழல் கட்டுண்ட இலக்கணம் (Context Sensitive Grammar), ஆக்கமுறை மாற்றிலக்கணம் (Generative Transfomational Grammar), கிளை சேர்ப்பு இலக்கணம் (Tree Adjoining Grammar), தொடரமைப்பு இலக்கணம் (Phrase Structure Grammar), தலைமை இயக்கத் தொடரமைப்பு இலக்கணம் (Head Driven Pharse Structurs Grammar (HPSG)), பாணினி இலக்கணம் (Panini

Grammar), வேற்றுமை இலக்கணம் (Case Grammar), சொல் செயல்பாட்டு இலக்கணம் (Lexical Functional Grammar (LFG)), பொதுமைப்படுத்தப்பட்ட தொடரமைப்பு இலக்கணம் (Generalized Phrase Strcture Grammar (GPSG) எனப்பல இலக்கண வடிவமைப்புகள் சொல்லமைப்பு ஆய்விற்காகவும் தொடரமைப்பு ஆய்விற்காகவும் பயன்படுத்தப்பட்டன.

தமிழ் போன்ற உருபனியல் அடிப்படையில் வளமுள்ள மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை உருவாக்கப்பட வேண்டுமானால் உருபனியல் பகுப்பாய்வு செம்மையாகச் செய்யப்பட்ட வேண்டும் எனக் கருதி உருபனியல் பகுப்பாய்விகள் (Morphological Analyzers) உருவாக்கப்பட்டன. உருபனியல் பகுப்பாய்வி இயந்திர மொழி பெயர்ப்பின் முக்கியமான ஒரு பாகமாகக் கருதப்பட்டு அதை உருவாக்கப் பல முயற்சிகள் மேற்கொள்ளப்பட்டன.

ஆங்கிலம் போன்ற மொழிகளுக்கு உருபனியல் வளத்தைக் காட்டிலும் தொடரியல் வளம் கூடுதலாக இருந்தன காரணமாகத் தொடரியல் பகுப்பாய்விற்கு முக்கியம் தரப்பட்டுத் தொடரியல் பகுத்துக் குறிப்பான்கள் (Syntactic Parsers) உருவாக்கப்பட்டன.

உருபனியல் பகுப்பாய்வும் தொடரியல் பகுப்பாய்வும் மூலமொழியில் செய்யப்பட்டு அதில் வரும் வெளியீடு இலக்கு மொழியில் மாற்றப்படும். இங்கு சிக்கலான செயல்பாடு என்னவென்றால் உருபனியல் பகுப்பாய்வின் வெளியீடு ஒன்றிற்கு மேற்பட்ட அர்த்தத்தைத் தரும் உருபனியல் பகுத்துக் குறித்தல்களையும் தொடரியல் பகுப்பாய்வின் வெளியீடு ஒன்றிற்கு மேற்பட்ட அர்த்தத்தைத் தரும் தொடரியல் பகுத்துக் குறித்தல்களைத் தருவதும் ஆகும். அதாவது மூலமொழியில் இருக்கும் ஒரு சொல்லுக்கு ஒன்றிற்கும் மேற்பட்ட உருபனியல் பகுத்துக் குறித்தல்களும் மூலமொழி வாக்கியத்திற்கு ஒன்றிற்கும் மேற்பட்ட தொடரியல் பகுத்துக் குறித்தல்களும் வருவது சாத்தியமாகும். ஒரு மூலமொழி வாக்கியத்திலிருந்து கிடைக்கும் ஒன்றிற்கு மேற்பட்ட பகுத்துக் குறித்தல்களில் ஒன்றுதான் சரியானது. அந்தச் சரியான ஒன்றுதான் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மாற்றப்படவேண்டும். எனவே பொருள் மயக்கம் நீக்கப்பட வேண்டிய

கட்டாயம் ஏற்படுகிறது. இதைச்சொல் மயக்க நீக்கம் (Words Sense disambiguation) என்பர். சொல் மயக்க நீக்கம் சூழல் அடிப்படையில் செய்யப்பட வேண்டியதாகும். இதற்காகப் பல புள்ளியியல் மாதிரிகள் இருக்கின்றன. இத்தகைய புள்ளியியல் மாதிரிகளைப் பயன்படுத்தி சொல் மயக்கம் நீக்கிச் சரியான தொடரியல் பகுத்துக் குறித்தலைப் பெற்று மூலமொழி இலக்கு மொழிக்கு மாற்றம் செய்யப்பட வேண்டும். மேற்கூறிய இயந்திர மொழிபெயர்ப்புச் சிக்கல்களையும் இயந்திர மொழி பெயர்ப்பின் எல்லைகளையும் கருத்தில் கொண்டு மொழி பெயர்ப்பிற்கான முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன.

மொழிபெயர்ப்பின் வெளிப்பாடு தெளிவாக இருக்க வேண்டும். சொற்பெயர்ப்பு மூல நூலின் சொற்களுக்கு இணையானதாகவும் எளிமையானதாகவும் சொற்கட்டு உடையதாகவும் இருக்க வேண்டும். மூல நூலில் உள்ளது போலவே கருத்துப் புலப்பாட்டுத் திறன் சிறப்பாக அமையவேண்டும். வாய்பாடு (Formula), சமன்பாடு (Equation), குறியீடு (symbol), சுருக்கக்குறியீடு அகியவற்றை அப்படியே பயன்படுத்தலாம். plan, scheme, project, programme என்பவற்றைத் 'திட்டம்' என்றே மொழிபெயக்கின்றனர். இச்சொற்களிடையே நுண்ணிய வேறுபாடுகள் உள்ளன. சொற்களுக்கிடையே உள்ள நுண்ணிய வேறுபாடுகளை அறிந்து மொழிபெயர்க்க வேண்டும். மூல மொழியிலிருந்து கருத்துக்களைப் பெயர்க்கும் போது சொற்பெருக்கம் நிகழ்கிறது. கடன் சொற்கள், கலப்புச் சொற்கள், கூட்டுச் சொற்கள், பழஞ்சொற்களுக்குப் புதுப் பொருள் எனப் பலவிதமாக இது அமைகின்றது. இதனால் மொழியமைப்பு விரிவடைகின்றது. புதுப் புது சொல்லகுகள் கிடைக்கின்றன. மொழியில் இலக்கண விரிவாக்கமும் நிகழ்கின்றது. மூல மொழியிலிருந்து குறிக்கோள் மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது குறிக்கோள் மொழியமைப்புக்கேற்ப மொழிபெயர்ப்பு அமைய வேண்டும். மொழிபெயர்ப்பு ஒரு மொழியிலிருக்கும் அறிவியல் கருத்துக்களை வேற்றுமொழிக்குப் பரிமாற்றம் செய்வதற்கு உதவுகின்றது. மொழிபெயர்ப்பு அறிவு வளர்ச்சிக்கும் மொழி வளர்ச்சிக்கும் மிகவும் உதவுகின்றது. கலைச்

சொல்லாக்கத்தில் மொழிபெயர்ப்புதான் பெரும் பங்கு வகிக்கின்றது. தமிழில் கலைச்சொற்கள் எல்லாம் ஆங்கில மொழிபெயர்ப்பாகும்.

எ.கா.

Telephone	:தொலைப்பேசி
Television	:தொலைக்காட்சி
University	:பல்கலைக்கழகம்
Photosynthesis	:ஒளிச்சேர்க்கை
Electron	:மின்னணு

இம்மொழிபெயர்ப்பு முற்காலத்தில் மனித முயற்சியால் மேற்கொள்ளப்பட்டு வந்தது. இதனால் காலவிரையமும் எளிதில் மொழிபெயர்க்க இயலாத நிலையும் ஏற்பட்டது. தற்காலத்தில் கணினியின் வரவால் மொழிபெயர்ப்புப் பணி முன்னேற்றம் அடைந்து கணினியைப் பயன்படுத்தி மொழிப்பெயர்ப்பு செய்யும் முறை மேற்கொள்ளப்பட்டு நடைப்பெற்று வருகின்றது.

2.2 கணிப்பொறி வழி மொழி பெயர்ப்பின் தேவை

ஒரு மொழியில் உள்ள தகவலை அதன் கருத்து மாறாமல் மற்றொரு மொழியில் மாற்றுவதையே மொழிபெயர்ப்பு என்கிறோம். இத்தகைய மொழி பெயர்ப்பில் தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகிறதோ அந்த மொழியின் இயல்புத் தன்மை மாறாது பாதுகாத்தல் வேண்டும். தகவல் எந்த மொழியில் உள்ளதோ அதை மூலமொழி (Source Language) என்கிறோம். தகவல் எந்த மொழிக்கு மாற்றம் செய்யப்படுகின்றதோ அந்த மொழியை இலக்குமொழி (Target Language) என்கிறோம். இத்தகைய மொழி பெயர்ப்புகளை மனித முயற்சியால் மட்டும் மேற்கொள்வது என்பது கடினமானப் பணியாகும்.

எனவே மொழிபெயர்ப்புப் பணிகளைக் கணிப்பொறி வழி மேற்கொள்வதற்கான கட்டாயம் ஏற்படுள்ளது. கணிப்பொறி வழி மொழிபெயர்ப்பின் தேவைகளும் கட்டாயங்களும் கீழேப் பட்டியலிடப்பட்டுள்ளன.

- உலகளவில் தகவல்கள் அன்றாடம் மலையெனக் குவிகின்றன. அவைகளை உடனுக்குடன் மொழிபெயர்த்தல் அவசியமாகிறது.
- கணிப்பொறி வழி மொழிபெயர்ப்பு மனித உழைப்பைக் குறைக்கிறது.
- குறைந்த நேரத்தில் நிறைய மொழி பெயர்ப்புப் பணிகளைச் செய்ய இயலுகிறது. இதனால் காலவிரயம் தவிர்க்கப்படுகிறது.
- தற்போதைய மொழி பெயர்ப்புப் பணிகளுக்கு ஏற்ற வகையில் மொழிப் பெயர்ப்பாளர்கள் இல்லை.
- மனித முயற்சியில் மொழிபெயர்கின்ற பொழுது மொழி பெயர்ப்பாளர்கள் மூலமொழியிலும் இலக்கு மொழியிலும் புலமை உள்ளவர்களிடம் கருத்துக்களைப் பெற்று அதை வழியமைப்பு மொழயில் எழுதி மொழிபெயர்ப்பு செய்யலாம்.

2.3 கணிப்பொறி வழி மொழிபெயர்ப்பின் வகைகள்

மொழிபெயர்ப்பில் கணிப்பொறி எவ்விதமாகப் பயன்படுத்தப்படுகின்றது என்பதன் அடிப்படையில் மொழிபெயர்ப்பைப் பின்வருமாறு பகுக்கலாம்:

1. கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு
   (CAMT – Computer Assisted Human Translation)
2. மனித உதவியின் கணிப்பொறி மொழிபெயர்ப்பு
   (HAMT-Human Assisted Machine Translation)
3. முற்றிலும் தானியங்கி மொழிபெயர்ப்பு
   (Automatic Machine Translation)

2.3.1 கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பு

இதில் கணிப்பொறி உதவியுடன் நடைபெறும் மனிதமொழி பெயர்ப்பில் மனித மொழிப் பெயர்ப்பாளர்களுக்கு முக்கியத்துவம் தரப்படுகின்றது. அவர்கள் மொழிப்பெயர்ப்பபுப் பணியில் ஈடுபடும் பொழுது ஒரு மொழியில் உள்ள சொற்களுக்கு நிகரான பிறமொழிச் சொற்களைத் தெரிந்துகொள்ளவும் அச்சொற்களின் பொருள் பயன்பாடு ஆகியவற்றை எளிமையாகத் தெரிந்துகொள்ளவும் கணிப்பொறி பயன்படுத்தப்படுகிறது. மேலும் ஒரு மொழியின் வாக்கிய அமைப்பு, தொடரமைப்பு போன்றவைகளுக்கு நிகரான பிறமொழி அமைப்புகளை விரைவாகத் தயார் நிலையில் பெறுவதற்கும் கணிப்பொறி உதவுகிறது. இவ்வகை மொழிபெயர்ப்பில் மொழிப்பெயர்பாளர்கள் தான் முதல் நிலையில் செயலாற்றுவர்; அவர்களுக்கு உதவும் நிலையில் கணிப்பொறி செயலாற்றும்.

2.3.2 மனித உதவியுடன் கணிப்பொறி மொழிப்பெயர்ப்பு

மனித உதவியுடன் நடைபெறும் இயந்திர மொழிப்பெயர்ப்பில் கணிப்பொறி முதன்மை நிலையை வகிக்கிறது. கணிப்பொறி மொழிப்பெயர்ப்புப் பணிகளைச் செய்கின்ற பொழுது முன்னர் பதிவு செய்யப்பட்டுள்ள நகலுக்கு மேலாக ஏதாவது செய்திகள் தேவைப்படின் அந்நிலையில் மனித உதவியை நாடுகிறது. மேற்கண்ட மூவகை மொழிப்பெயர்ப்பு வகைகளில் மனித உதவியுடன் செயல்படும் இயந்திர மொழி பெயர்ப்பு தான் நடைமுறையில் பரவலாக உள்ளது.

2.3.3 முற்றிலும் தானியங்கி மொழிபெயர்ப்பு

முற்றிலும் இயந்திர மொழிபெயர்ப்பு என்பது தற்போதைய நிலையில் சாத்தியமானதல்ல இதற்கு எண்ணற்ற மொழியில் செயல்பாடுகளும் மொழிபெயர்ப்பு முறைமைகளும் வடிவமைக்கப்பட்ட வேண்டி உள்ளது.

இருப்பினும் கணிப்பொறி வழி மொழிபெயர்ப்பில் முற்றிலும் தானியங்கப்படுத்தும் முயற்சி இலக்காகக் கொள்ளப்பட்டுள்ளது.

2.4 கணிப்பொறி வழி மொழிப்பெயர்ப்பின் சில முயற்சிகள்

கணிப்பொறி வழி மொழி பெயர்ப்பினைத் தானியங்கி மொழிப்பெயர்ப்பு (Automatic Translation) என்றும் இயந்திர மொழிபெயர்ப்பு (Machine Translation) என்றும் கணிப்பொறி உதவியுடன் நடைபெறும் மொழிபெயர்ப்பு என்றும் கூறலாம். கணிப்பொறியின் வேகம், நினைவகத்திறன், பதிவு செய்துள்ள செய்திகளைத் தேவையான பொழுது தேவையான வடிவமைப்பில் மிக விரைவாக மீளப்பெறக்கூடிய வசதி (Retrievel System) போன்றவைக் கணிப்பொறி வழி மொழி பெயர்ப்புப் பணியை ஊக்குவிக்க உதவின.

கணிப்பொறி வழி மொழிபெயர்ப்புப் பணி 1950-ஆம் ஆண்டு வாரன் வீவர் (Warren Waiver) என்ற அறிஞரால் தொடங்கப்பட்டது. இதைத் தொடர்ந்து பூபார் ஹில்லல் என்ற அறிஞர் கணிப்பொறி வழி மொழிபெயர்ப்புப் பணியைத் தொடங்கினார். 1954-ஆம் ஆண்டு ஜார்ஜ் டவுன் பல்கலைக்கழகம் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியைத் தொடங்கியது. இதைத் தொடர்ந்து 1956-ஆம் ஆண்டு மாஸ்கோவில் பிரெஞ்சு மொழியிலிருந்து ரஷ்ய மொழிக்குக் கணிப்பொறி வழி மொழிப்பெயர்ப்பு செய்யும் பணி நடைபெற்றது. இப்பணியை மேல் சங் (Mel Chunk) என்ற அறிஞர் மேற்கொண்டார்.

1973-ஆம் ஆண்டு அமெரிக்காவில் உள்ள Logos Incoration Limited என்ற நிறுவனம் ஒரு மென்பொருளை வடிவமைத்து. இம்மென் பொருளைக் கொண்டு ஆங்கிலத்தில் அமைந்துள்ள செய்திகளை வியட்நாம் மொழியில் மொழிப்பெயர்ப்பு செய்தது. ஏறக்குறைய இதே காலக்கட்டத்தில் Tertils என்ற நிறுவனம் TUTUS என்ற மென்பொருளை வடிவமைத்தது. இம்மென் பொருளைக் கொண்டு ஆங்கிலம், ஸ்பானிஷ், பிரெஞ்சு மற்றும் ஜெர்மன் ஆகிய மொழியில் உள்ள

செய்திகள் ஒன்றிலிருந்து மற்றவைகளுக்கு மொழிமாற்றம் செய்யப்பட்டது. இதைத் தொடர்ந்து CUTD-Chinese University Translator) என்ற மென்பொருள் உருவாக்கப்பட்டது. இதைக்கொண்டு சீனமொழியில் உள்ள செய்தித்தாள்கள் ஆங்கில மொழிக்கு மொழிமாற்றம் செய்யப்பட்டன.

1974-ஆம் ஆண்டு SYSTRAN என்ற மென்பொருள் உருவாக்கப்பட்டது. SYSTRAN என்பதன் விரிவு System Translation என்பதாகும். இம்மென்பொருள் ரஷ்ய மொழியிலிருந்து பிரஞ்சு மொழிக்கு மொழி மாற்றம் செய்யப்பயன்பட்டது. இதற்கானத் திட்டத்தை P.Toma என்பவர் மேற்கொண்டார். வணிக ரீதியில் முதன்முதலில் வெளிப்யிடப்பட்ட கணிப்பொறி வழி மொழிபெயர்ப்புக்கான மென்பொருள் இதுவேயாகும்.

1975-ஆம் ஆண்டு TAUM என்ற மென்பொருள் உருவாக்கப்பட்டது. இம்மென் பொருள் ரஷ்ய மொழியில் மாற்றம் செய்யப்பட்டது. இதே காலக்கட்டத்தில் GETA என்ற மென்பொருள் ஆங்கில மொழியிலிருந்து பிரெஞ்சு மொழிக்கு வழியமைக்க உருவாக்கப்பட்டது. இது போன்று உலகளாவிய நிலையில் ஆங்காங்கே சோதனைக்கட்ட முயற்சியாலும் வணிக நோக்கத்தாலும் கணிப்பொறி வழி மொழிபெயர்க்க மென்பொருள்களை உருவாக்கும் திட்டங்கள் தொடர்ந்து நடைபெற்று வருகின்றன.

1985-ல் தமிழ்ப்பல்கலைக் கழகத்தில் TUMTS (Tamil University Mechine Translation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. ரஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழி பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப் பல்கலைக்கழகக் கணிப்பொறி அறிவியல் துறை, மொழியியல் துறை, மொழிப்பெயர்ப்புத்துறை ஆகியவைகளின் கூட்டு முயற்சியில் நடைமுறைப்படுத்தப்பட்டது. இதற்குப்பின்னால் கான்பூரில் உள்ள இந்திய தொழில்நுட்ப நிறுவனம் Indian Institute of Technology இந்தியிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து இந்திக்கும் மொழிபெயர்ப்பு செய்வதற்கும்

இந்திய மொழிகளை இந்திக்கும் இந்தியை பிற இந்திய மொழிகளுக்கும் மொழிபெயர்ப்பதற்கான ஆரம்ப கட்ட வேலைகளை மேற்கொண்டு செயலாற்றியது. அனுசாரகா (*anusaraka*) என்ற இயந்திர மொழிபெயர்ப்பு மென்பொருள் இதன் விளைவாகும்.

2.5 கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள்

ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழி மாற்றம் செய்வதற்கான மென்பொருள்கள் பல உருவாக்கப்பட்டு வருகின்றன. பெரும்பாலும் இத்ததைய கணிப்பொறி வழி மொழி பெயர்ப்புப் பணிகளில் எல்லைகள் இரண்டாக வரையறுத்து உள்ளன.

1. மொழி என்பது பல வட்டாரங்களில் மக்கள் பேசும் மற்றும் எழுதும் அமைப்புகளின் தொகுப்பாகும். ஒரு வட்டாரத்தில் பயன்படுத்தக்கூடிய சொற்கள் பிற வட்டாரங்களில் பயன்படுத்தக்கூடிய உறுதியில்லை. தமிழ் போன்ற மொழிகளில் மொழி நடை இலக்கியமொழி, பேச்சுமொழி என இருவகையாகப் பிரிக்கப்படுகிறது. பேச்சுமொழியைக் கணிப் பொறியாக்கம் செய்வது என்பது சுலபமான பணியன்று. எனவே, மொழிப்பெயர்ப்புப் பணியில் இலக்கிய மொழிகள் மட்டுமே எடுத்துக்கொள்ளப்படுகின்றன.

2. செய்யுள், கவிதை போன்றவைகளில் பொருளுக்குத் தரும் முக்கியதுவத்தைவிட கருத்தைப் புலப்படுத்தும் உத்தி, உணர்ச்சி, வேகம் போன்ற கூறுகளுக்கு முக்கியத்துவம் தரப்படுகின்றன. இவைகள் மனித உளவியலைப் பிரதிப்பலிப்பன. இவ்வுளவியல் கூறுகளைக் கணிப்பொறி வழி வெளிப்படுத்துவது என்பது இயலாத ஒன்றாகும். இதேபோல் இலக்கியங்களிலும் பொருளைவிட உளவியல் கூறுகளுக்கு தான் முக்கியத்துவம் தரப்படுகின்றது. எனவே, செய்யுள்கள், கவிதைகள்,

பாடல்கள் மற்றும் இலக்கியச் செய்திகள் ஆகியவைகளைக் கணிப்பொறி வழி மொழிபெயர்ப்பதற்குத் தொடக்கநிலையில் பயிற்சி எடுத்துக்கொள்ளப்படவில்லை. அறிவியல் தொழில்நுட்ப சார்ந்த படைப்புகளில் பொருள்களுக்கே முக்கியத்துவம் தரப்படுகிறது. இங்கு கோபம், வீரம், தாபம், மகிழ்ச்சி போன்ற உணர்வுகளுக்கு முக்கியத்துவம் தரப்படுவதில்லை. எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணிகளுக்குப் பொதுவாக அறிவியல் மற்றும் தொழில்நுட்பத்தைச் சார்ந்த படைப்புகளே எடுத்துக்கொள்ளப்படுகின்றன.

இதேப் போன்று மொழிகளின் தன்மையைப் பொறுத்துக் கணிப்பொறி மொழிபெயர்ப்பினை இருபெரும் பிரிவுகளாகப் பிரிக்கலாம்:

1. இருதொடர்புடைய மொழிகளுக்கிடையே அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Interlingual Method of Machine Translation): இலக்கண அமைப்பிலும் பொருள் அமைப்பிலும் தொடர்பும் நெருக்கமும் உடைய இருமொழிகளுக்கிடையே செய்யப்படும் கணிப்பொறி மொழிப்பெயர்ப்பு இவ்வகையைச் சார்ந்தாகும்.

2. இடைநிலை மொழி அடிப்படையில் அமைந்த கணிப்பொறி மொழிப்பெயர்ப்பு (Intermediate Language Based Machine Translation) இலக்கணம் மற்றும் மொழியமைப்பில் முற்றிலும் வேறுபட்ட இருமொழிகளுக்கிடையே கணிப்பொறி வழி மொழிபெயர்ப்பை நடைமுறைப்படுத்த இம்முறை உதவுகிறது. இம்முறையில் மூலமொழியில் உள்ள சொற்றொடர்கள் இடைநிலை மொழி வழி இலக்கு மொழிக்கு மாற்றம் செய்யப்படுகின்றது.

2.6 கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும்

மொழிபெயர்ப்புப் பணியைக் கணிப்பொறி வழி மேற்கொள்கின்ற பொழுது கணிப்பொறியின் சிறப்புத் தன்மை, துல்லியம், வேகம், நினைவகக் கொள்திறன், தகவல் தளம் உருவாக்கும் வசதிகள் போன்றவை மொழிபெயர்ப்புத் திறனை நிர்ணயம் செய்யும். ஒரு மொழியிலிருந்து பிற மொழிக்கு மொழிபெயர்ப்பு செய்யும் போது பன்மொழித் திறமை அவசியம் என்ற நிலை இல்லை. அந்தந்த மொழியின் வல்லுநர்களின் மொழித்திறன் கணிப்பொறி வழியமைப்பாக வழியமைக்கப்படுகிறது. மனித மொழிபெயர்ப்புப் பணிக்குக் கூட ஒருவருக்கு இருமொழியிலும் திறமை இருக்கவேண்டும் என்ற அவசியம் இல்லை.

ஒரு மொழியின் வளம், தன்மை ஆகியவற்றைக் கணிப்பொறிக்கு ஏற்றாற்போல் கணிதப்படுத்துவது என்பது எளிமையான பணி அல்ல. எனவே, கணிப்பொறி வழி மொழி பெயர்ப்புத் திட்டங்கள் சுமார் 50 ஆண்டுகளுக்கு முன்னர் தொடங்கப்பட்டிருப்பினும் இன்னும் முழுமைபெறவில்லை. மேலும் கணிப்பொறி அறிவியலில் தொழில் நுட்ப வளர்ச்சி கணிப்பொறி வழி மொழிப்பெயர்ப்பு பணிக்கு மகிவும் பயனுள்ளதாக அமைகின்றது. இத்தகைய பணிகளின் நிறை குறைகளைக் கீழேக் காணலாம்.

மொழிக்கு இடைடே உள்ள சொல் வளம் இலக்கணம் கருத்தாடல் போன்றவைகளின் ஒற்றுமைகளை ஆராய வேண்டும். இவைகளில் மொழிக்கு இடையே உள்ள ஒற்றுமை அதிகரிக்க அதிகரிக்க மொழிபெயர்ப்புப் பணியின் சிக்கல்கள் குறையும்.

உருப்பனியல் அமைப்பை பொறுத்தவரையில் மொழிகளின் பண்புகள் இருவேறு திசைகளில் மாறுகின்றன:

அ. ஒரு சொல்லில் எத்தனை உருபன்கள் உள்ளன என்பதைப் பொறுத்து.

ஆ. ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இருந்தாலும் அதற்கு மேலே இருந்தாலும் அவைகளைத் துல்லியமாகப் பிரிக்க இயலும் என்பதைப் பொறுத்து.

ஒரு சொல்லில் ஒரே ஒரு உருபன் மட்டும் இந்தால் அந்த மொழியைத் தனிநிலை உருபன்கட்டு மொழி (isolating Language) மொழி என்கிறோம். ஒரு மொழியில் உள்ள சொற்களில் ஒன்றுக்கு மேற்பட்ட உருபன்கள் இருந்தால் அதை பன்நிலை உருபன் கட்டுமொழி (Poly Synthetic) என்கிறோம். சிபேரியன் மொழி பன்நிலை உருபன் கட்டுமொழிக்குச் சிறந்த உதாரணமாகும். வியட்நாமிஸ் மற்றும் கேண்டுனஸ் ஆகிய மொழிகள் தனிநிலை உருபன்கட்டு (isolating) மொழிக்குச் சிறந்த உதாணரங்களகும். எனவே இத்தகைய மொழிக்குச் மொழி மாற்றம் செய்ய கணிப்பொறி வழி திட்டங்கள் வகுக்கின்ற பொழுது மொழியின் உருபனியல் பற்றிய விரிவான தகவல் அவசியம் ஆகும்.

மொழியின் தொடரியல் அமைப்புகள் (Syntactic Structure) மூன்று வகைகளாய் பிரிக்கப்படுகின்றன.

1. SVO அமைப்பு

ஜெர்மன், பிரஞ்சு, ஆங்கிலம் போன்றவை SVO அமைப்பில் உள்ள மொழிகளாகும்.

2. SOV அமைப்பு

தமிழ், இந்தி, ஜப்பான் மொழிகள் SOV அமைப்பில் உள்ள மொழிகளாகும்.

3. VSO அமைப்பு

இர்ரீஸ், அரேபி, ஷிபிரோ போன்ற மொழிகள் VSO அமைப்பில் அமைந்துள்ளன.

SVO அமைப்பில் உள்ள மொழிகளில் முன்னுருப்புகள் (Preposition) பயன்படுத்தப்படுகிறது. எனவே கணிப்பொறி வழி மொழிபெயர்ப்புத்

திட்டங்களுக்கு மொழிகளுக்கு இடையே உள்ள தொடரியல் அமைப்பு வகைகள், உறவுகள் போன்றவை ஆராயப்பட வேண்டும்.

இராமனின் வீடு என்ற தொடரில் *வீடு* என்பது தலைச்சொல் (Head Word) ஆகும். *இராமன்* என்பது சார்புச் சொல் (Dependent Word) ஆகும். தலைச்சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையில் உள்ள உறவை *-இன்* என்ற ஒட்டு விளக்குகிறது. இது போன்று தலைச் சொல்லுக்கும் சார்புச் சொல்லுக்கும் இடையிலான முறையைப் பொறுத்து மொழி தலைக்குறித்தல் மொழி, சார்புக்குறிதல் மொழி எனப் பிரிக்கப்படும். சார்பு குறித்தல் மொழியில் ஒட்டு சார்புச் சொல்லுடனும் தலைக்குறித்தல் மொழியில் ஒட்டு தலைச் செல்லுடனும் தரப்படுகின்றன. ஆங்கில மொழி தலைக்குறித்தல் மொழியைச் சார்ந்ததாகும். ஹங்கேரிய மொழி சார்புக்குறித்தல் மொழியைச் சார்ந்ததாகும். பின்வரும் எடுத்துக்காட்டு விளக்கும்.

English	:	The Man's house
Hungarion	:	The man house his
House	:	Head word (தலைச்செல்)
The Man	:	Dependent word (சார்புச்சொல்)
-s	:	Affix (ஒட்டு)

எனவே கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியை மேற்கொள்கின்ற போது தலைச்சொல்லுக்கும், சார்பு சொல்லுக்கும் இடையே உறவு முறைகள் எவ்வாறு குறிக்கப் பெறுகின்றன என்பது பற்றியெல்லாம் ஆராயவேண்டும்.

மூலமொழியில் உள்ள ஒரு சொல்லின் வகைப்பாடும் இலக்கு மொழியிலுள்ள அதற்கு நிகரான சொல்லின் வகைப்பாடும் ஒன்றாக இருக்கவேண்டியத் தேவையில்லை. எடுத்துக்காட்டாக ஒரு மொழியிலுள்ள வினைச்சொல் பிறமொழிக்கு மொழி பெயர்ப்பு செய்யப்படும் போது

மொழிப்பெயர்ப்புச் சொல் வினைச்சொல்லாக இருக்கவேண்டிய தேவையில்லை. இவைகள் எல்லாம் மொழிபெயர்ப்புப் பணியில் உள்ள சிக்கல்களாகும்.

ஒவ்வொரு மொழியிலும் கருத்தாடல்களுக்கானச் சொல் விளக்கமானதாகவோ கருக்கமானதாகவோ இருக்கலாம். எடுத்துக்கட்டாக *சகோதரன்* என்ற உறவைக் குறித்து ஆங்கிலத்தில் *brother* என்ற சொல் உள்ளது. மூத்த சகோதரனை ஆங்கிலத்தில் *elder brother* என்று சொல்லுகிறோம். இவ்வாறு வயது மூத்தவர் இளையவர் என்பதைக் குறிக்க *elder*, *younger* என்ற அடைச் சொற்கள் பயன்படுத்தப்படுகின்றன. ஆனால் தமிழில் *அண்ணன்* என்றும் *தம்பி* என்றும் தனித்தனிச்சொற்கள் உள்ளன. *சாகோதரன்* என்ற உறவை மேலும் வகைப்படுத்த முதியவர், இளையவர் என்ற நிலையில் *அண்ணன்*, *தம்பி* என்ற இருச்சொற்கள் உள்ளன. ஆனால் ஆங்கிலத்தில் இவ்வகையிலான வேறுபாடு என்பது இல்லாதால் உறவை வேறுப்படுத்த *elder*, *younger* என்ற சொற்களைப் பயன்படுத்த நேர்கிறது.

மூலமொழியில் உள்ள செய்திகளில் கலாசாரப் பின்னணி மொழிப்பெயர்ப்பிற்கு ஒரு முக்கியமான காரணியாகும். அதே கலாச்சாரப் பண்பு இலக்கு மொழியின் கலாச்சாரப் பண்புகளுக்கு ஏற்ப மாற்றித்தரத்தக்க அளவு இருத்தல்வேண்டும். இருமொழிகளுக்கு இடையே உள்ள கலாச்சார தொடர்புகள் மொழிபெயர்ப்பு செய்ய இயலுமா? இயலாதா? என்பதைக் தெளிவுபடுத்த உதவும். மேலும் சபீர்-ஊர்ஃப் கொள்கையின் படி ஒருவர் பேசும் மொழி சிந்திக்கும் திறனைக் கூட்டுவதாகவோ குறைப்பதாகவோ உள்ளது. எனவே மொழிபெயர்ப்பில் இத்தகைய உளவியல் அடிப்படையிலானச் சிக்கல்கள் உள்ளன.

தொடரியல் பகுப்பாய்வி பல பணிகளுக்காக உருவாக்கப்படுகின்றன. ஆனால் இயந்திர மொழிபெயர்ப்பிற்காகப் பயன்படுத்தப்படும் பகுப்பாய்வு மற்ற பகுப்பாய்வை விட சிறிது வேறுபடுகிறது. இங்குத் தொடரின் அல்லது

வாக்கியத்தின் பொருளை அடிப்படையாகக் கொண்டு பகுப்பாய்வு செய்யப்படுகிறது. எனவே இயந்திர மொழிபெயர்ப்பிற்கு உருவாக்கப்படும் பகுப்பாய்வு தொடரியல் அமைப்பை மட்டும் சாராது பொருளைக் சார்ந்ததாகவும் அமைகிறது. இத்தகைய பகுப்பாய்வுகளை உருவாக்க அறிவு அடிப்படை (Knowledge Base) தேவைப்படுகின்றது.

இயந்திரம் வழி மொழிபெயர்ப்பு செய்கின்ற பொழுது ஏதேனும் சிக்கல்கள் இருந்தாலோ அல்லது கணிப்பொறிக்குக் கூடுதல் கதவல் தேவைப்பட்டாலோ மனித உதவியை எதிர்நோக்குகின்றது. இத்தகைய சூழல்களில் சிக்கலைப் புரிந்துக்கொண்டு அதற்கேற்றத் தீர்வு உரிய முறையில் வழங்கப்பட்டவேண்டும். எனவே இயந்திர மொழிப்பெயர்ப்புத் திட்டத்தை இயந்திர மனித உரையாடலுக்கு ஏற்றாற்போல் வடிவமைக்கவேண்டியது அவசியம் ஆகும். பெரும்பாலும் படைப்பாற்றல் திறனைச் செய்திகளில் வெளிப்ப்படுத்தும் பொழுது கணிப்பொறிக்குக் கூடுதல் செய்திகள் தேவைப்படுகிறன்றன.

பெரும்பாலான கணிப்பொறி வழி மொழிபெயர்ப்புகளிலும் உள்ளீடு செய்யும் செய்திகளை முன்திருத்தம் செய்யும் பொழுது மூலமொழியின் தன்மையும் இலக்கு மொழியின் தன்மையும் கருத்தில் கொள்ளப்படுகின்றன. மூலமொழியில் உள்ள செய்தியை முதலில் இலக்கு மொழிக்கு மொழிபெயர்ப்பது, பின்னர் மொழி பெயர்ப்பைப் படித்தறிந்து அது சரியான கருத்தை வெளிப்படுத்துகின்றதா என்பதை ஆராய்வது, அப்படி இல்லையெனில் மூலமொழி செய்தியை முன்திருத்தம் செய்து மறுபடியும் மொழிபெயர்ப்பு செய்வது, சரியான கருத்து வெளிப்படுவதுவரை இத்தகைய ஆய்வை மீண்டும் மேற்கொண்டு மொழி பெயர்ப்பிற்கு முன்னால் முன்திருத்தப் பணிக்கு முக்கியத்துவம் அளித்து தேவையான திருத்தங்களைச் செம்மையாக மேற்கொள்வது. இம்முறை கணிப்பொறியின் நேரத்தை மிச்சப்படுத்துகிறது; மொழிபெயர்ப்பின் வேகத்தை அதிகப்படுத்துகிறது. மேலும் ஒரு மொழியிலிருந்து பல மொழிகளுக்கு மொழிபெயர்ப்பு செய்ய வேண்டிய சூழல்களில் முன்திருத்தப்பணித் திறமையானதாக அமைகிறது.

மொழிபெயர்ப்புப் பணிகளுக்காக உருவாக்கப்படுகின்ற அகராதிகள் சாதாரண முறையில் அல்லாது பல கூடுதல் செய்திகளைக் கொண்டதாக இருத்தல் வேண்டும். பிறவகை அகராதிகளை விட துறை சார்ந்த அகராதிகள் மொழிபெயர்ப்புப் பணியை எளிமைப்படுத்துகிறது. மூலமொழியில் உள்ள செய்திகளைப் பொறுத்து அது தொடர்புடைய அகராதியைப் பயன்படுத்திக்கொள்ளலாம்.

மொழிபெயர்ப்புக்காக உருவாக்கப்படும் அகராதியில் சொற்கள், தொடர்கள், மரபுத் தொடர்கள் மேலும் அடிக்கடி பயன்படுத்தக்கூடிய வாக்கியங்கள் போன்றவைகள் உரிய விளக்கத்துடன் தரப்படவேண்டும்.

கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணி என்பது ஒரே மூச்சில் ஒட்டு மொத்தமாகச் செய்யக்கூடிய பணியாகும். முதலில் கணிப்பொறி வழியமைப்பிற்கான மென்பொருள் வடிவமைக்கப்படவேண்டும். அதன் செயல் திறனை அறியவேண்டும். பணியின் தன்மையை மதிப்பிடவேண்டும். இதற்கு ஏற்றார் போல் தேவைப்படின் மென்பொருளைக் கூடுதல் வசதிகளுடன் மேம்படுத்தப்பட வேண்டும். இம்முறைகளை மனிதனுக்கு நிறைவழிக்கும் கொள்திறன் கிடைக்கின்ற வரையில் திரும்பத் திரும்பச் செய்ய வேண்டும்.

கணிப்பொறி வழி மொழிபெயர்ப்புக் கடுமையானப் பணியாகும். இதற்கு மனிதமொழி பெயர்ப்பில் மேற்கொள்ளப்படும் திட்ட மாதிரிகளும் (Model) வழிமுறைகளும் அவசியமாகும். கணிப்பொறி வழி மொழிபெயர்ப்பு எண்ணற்ற முயற்சிகளால் உருவாக்கப்படுகின்ற ஆக்கப்பூர்வமான பணியாகும். இதற்கு அறிவியல் திறன் மட்டுமல்லாது மொழிபெயர்ப்பிற்குத் தேவையான கலைத்திறனும் அவசியமாகும்.

தற்பொழுது உலகளவில் நடைபெற்றுள்ள மொழிபெயர்ப்புப் பணிகள் மனதுக்கு நிறைவழிக்கக்கூடிய வகையில் உள்ளன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்கு உதவுகின்றன. இவைகள் ஓரளவுக்கு மொழிபெயர்ப்பிற்குப் போதுமானதாகும்.

மனிதனின் படைப்பாற்றல் கற்பனைத்திறன், உளவியல்பாங்கு போன்ற செய்திகளால் பொதிந்துள்ளதால் கணிப்பொறி வழி மொழிப்பெயர்ப்புப் பணியில் இத்தகையக் கூறுகளைச் சேர்க்கவேண்டி உள்ளது. அதற்கு மனித நுண்ணிறிவுடன் கூடிய கணிப்பொறிகள் உருவாக்கப்படுகின்றன.

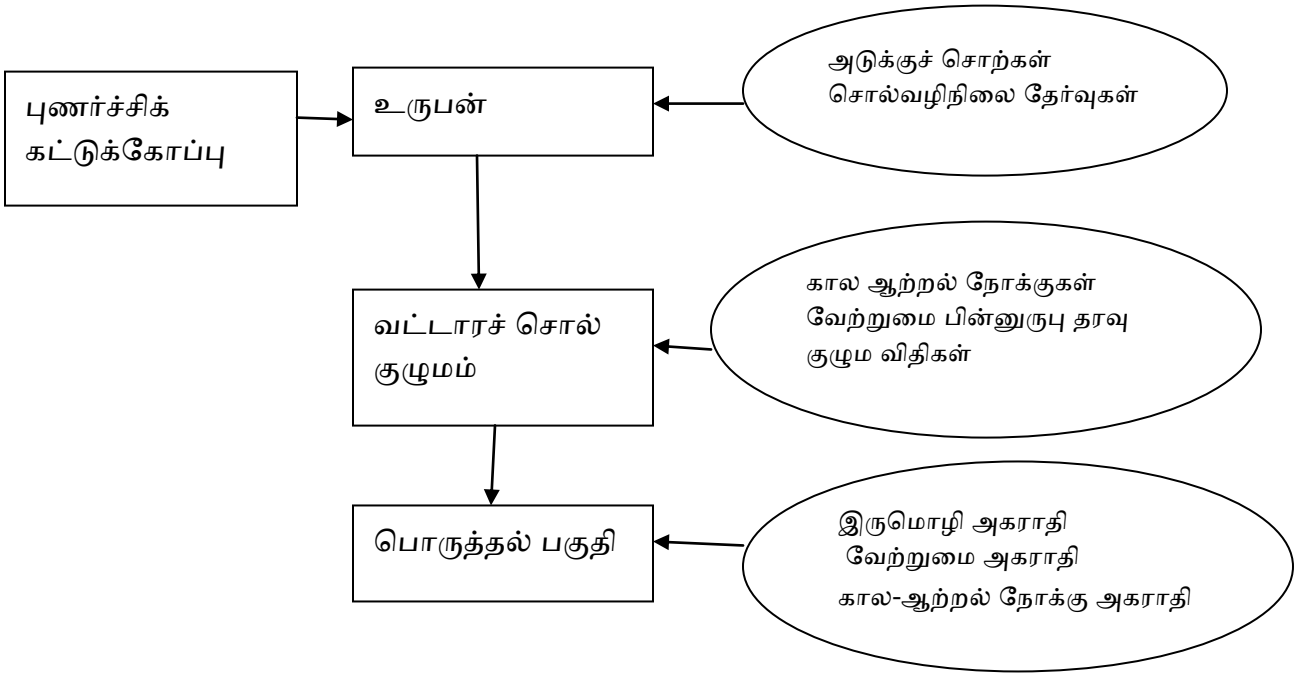
2.7 இந்தியாவில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சி

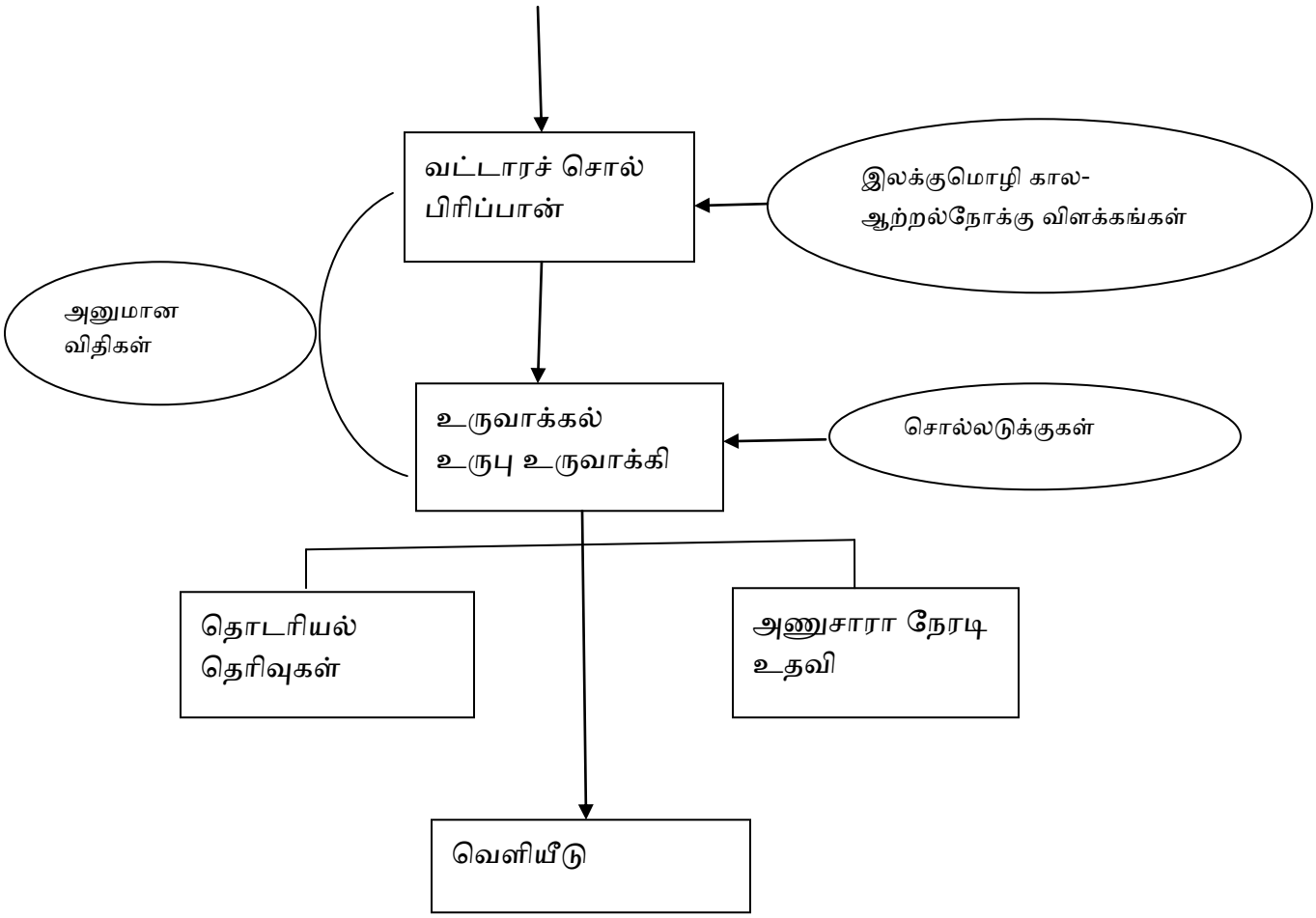
ஆரம்பக்கட்டத்தில் இந்தியாவில் இயந்திர மொழிபெயர்ப்புகள் இதற்கான சாத்தியத்தை வெளிப்ப்படுத்தும் பொம்மை ஒழுங்குமுறைகளாக (Toy Systems) இருந்தன. 1978 வரை இயந்திர மொழிபெயர்ப்பிற்கான எந்த முயற்சியும் எடுக்கப்படவில்லை. 1978-இல் ஐ.ஐ.டி.கான்பூரில் உள்ள மின்னியல் தொழில் நுட்பத் துறையின் முயற்சியால் மொழியியல் அடிப்படையாலன தகவல் ஒழுங்குமுறைகள் பற்றிய தேசியக் கருத்தரங்கு (National Symbosium on Linguistic Based Information System) முதன் முதலில் நடைபெற்றது. இக்கருத்தரங்கு ஆராய்ச்சியாளர்களிடையேயும் அரசாங்கத்திலும் தொழில்நுட்ப துறைகளிலும் இயந்திர மொழிபெயர்ப்பு குறித்த விளிப்புணர்ச்சியை ஏற்படுத்தியது. இருப்பினும் இயந்திர மொழிபெயர்ப்பு 80-களின் ஆரம்பகட்டத்தில் இந்திய மொழிகளுக்குச் செல்லாய்விகளும் பிற கணிபொறி வசதிகளும் வரத் தொடங்கிய பின்னர்தான் தொடங்கப்பட்டது. ஐ.ஐ.டி.கான்பூர் GIST என்ற தொழில் நுட்பத்தின் உருவாக்கத்திற்கு முயன்றது. இது இயந்திர மொழிபெயர்பிற்கு விழிப்புணர்ச்சி ஏற்படுத்துவதுடன். இதில் ஆய்வு மற்றும் வளர்ச்சி குறித்த பல திட்டங்களை ஊக்குவித்தது. இந்தியாவில் பல மையங்களில் பலவிதமான முயற்சிகள் மேற்கொள்ளப்பட்டாலும் மூன்றும் அணுகுமுறைகளைப் பட்டியலிட இயலும்.

1. இடமெழி அணுகுமுறை (Interlingual Approach)
2. நேரடி சொல்சார் மாற்றம் (Direct lexical Transfer Approach)
3. கணிப்பொறி உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிகள் (Machine Aided Translation Effort)

2.7.1 இடைமொழி அணுகுமுறை (Interlingual approach)

சமஸ்கிருதத்தை இடைப்பட்ட அடிப்படை மொழியாகக் கொண்டு இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பு செய்யும் இடைமொழி ஆய்வு அணுகுமுறை சின்ஹா என்பவரால் 1984-இல் முயற்சிக்கப்பட்டு 1989-இல் விரிவாக்கப்பட்டது. இது மூல மொழியின் ஆய்வுக்குக் காரக அடிப்படையிலான அக உருப்படுத்தத்தை முன்மொழிந்தது. 1986-1988-களில் எல்லைகுட்பட்ட சொற்களைக் கொண்ட மிக எளிய வாக்கியங்களை இந்தியிலிருந்து தெலுங்கிற்கு மொழிபெயர்க்கும் மூலமுன்மாதிரி (prototype) சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் உருவாக்கப்பட்டது. இது காரக அடிப்படையிலான கருத்துவரைபடத்தை அக உருப்படுத்த்திற்கும் பொருண்மை மயக்கங்களுக்கும் பயன்படுத்தியது. ஒரு தனிநிலை தொடரியல் இருப்பானில் உள்ள மிகக்கூடுதலாகத் தொடர்புள்ள சொற்கள் ஒரு சொல் குழுவாகப் பயன்படுத்தப்பட்டது. மூல மொழியில் உள்ள உரை உருவாக்கி இச்சொற் குழுமங்களிலிருந்து பொருத்தத்தைச் சரியான உருபனியல் உருவாக்கத்திற்குப் பயன்படுத்தியது.


புணர்ச்சிக் கட்டுக்கோப்பு
உருபன்
அடுக்குச் சொற்கள்
சொல்வழிநிலை தேர்வுகள்
வட்டாரச் சொல் குழுமம்
கால ஆற்றல் நோக்குகள்
வேற்றுமை பின்னுருபு தரவு
குழும விதிகள்
பொருத்தல் பகுதி
இருமொழி அகராதி
வேற்றுமை அகராதி
கால-ஆற்றல் நோக்கு அகராதி



இடைமொழி அடிப்படையிலான இயந்திர மொழிபெயரிப்பு அதற்குப்பின் முயற்சிக்கப்படவில்லை. இவ்வணுகுமுறை இந்திய மொழிகளின் பன்மைத்தன்மைக்குக் கூடுதல் அனுகூலமாய் இருந்தாலும் அக உருப்படுத்ததிலிருந்து புற அமைப்பை பெறுவது கடினமான செயல்பாடாக அமைந்து குறையுடையதாய் இருந்தது. இவ்வணுகுமுறை மூலமொழியின் சீரான புரிதலை முன்கருதலாகக் கொண்டதால் இன்றைய ஆய்வுநிலையில் ஏற்றதாக அமையவில்லை. மேலும் அமைப்பு அடிப்படையில் அண்மைப்படுகின்ற இந்திய மொழிகளுக்கிடையில் மொழிபெயர்ப்பின் போது நேரிடையான சொல் பதிலீடு செய்வது எளிமையாக அமைந்தது. இதன் காரணமாக இவர்களால் அனுசாரகா என்ற அணுகுமுறை உருவாக்கப்பட்டது.

இந்திய மொழிகளுக்கிடையே உள்ள அமைப்பு ஒற்றுமை காரணமாக மூலமொழியிலிருந்து இலக்கு மொழிக்குப் பல பொருண்மை மயக்கங்களைக் கொண்டு செல்ல இயலும் என்ற காரணத்தால் இந்த இடைமொழி அணுகுமுறை இந்திய மொழிகளுக்கிடையே மொழிபெயர்ப்பிற்கு நல்லதொரு அணுகுமுறையாக அமையும். இருப்பினும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கும் இந்திய மொழிகளிலிருந்து ஆங்கில மொழிகளுக்கும் மொழிபெயர்க்க இடைமொழி அணுகுமுறை பொருத்தமான அணுகுமுறை அல்ல.

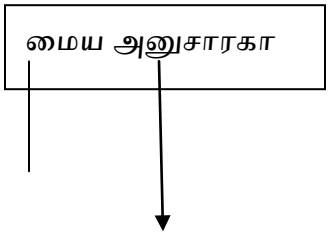
2.7.2 சொல்நிலையில் மாற்றத்தைப் பயன்படுத்திய நேரடி அணுகுமுறை (அனுசாரக் திட்டம்)

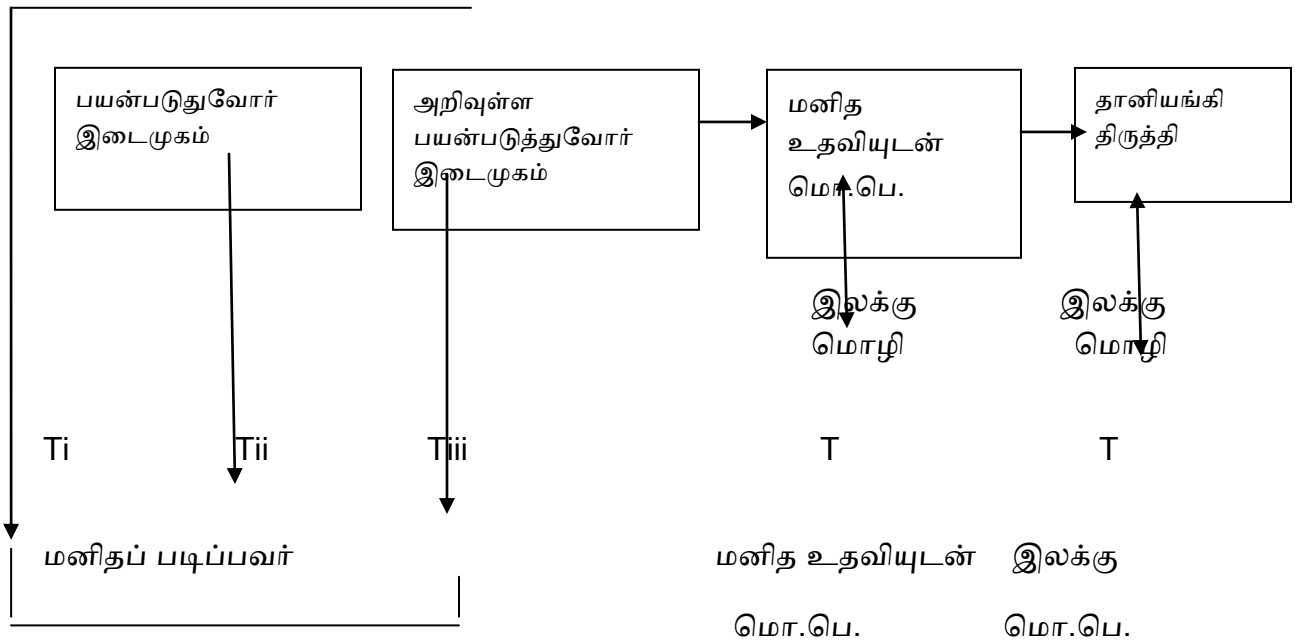
சைத்தன்யா மற்றும் ராஜீவ்சங்கல் என்பவர்களால் மூலமொழியிலிருந்து இலக்கு மொழிக்கு நேரடி சொல்சார் மாற்ற அணுகுமுறை உருவாக்கப்பட்டது. இங்கு மூல மொழிக்கும் இலக்கு மொழிக்கும் இடையில் உள்ள வாக்கிய அமைப்பின் பொதுமைப் பண்புகள், இரண்டும் இந்திய மொழிகள் என்ற காரணத்தால் கருதப்பட்டு முழு அளவிலும் பயன்படுத்த்ப்பட்டுள்ளது. இவ்வணுகுமுறையின் படி ஒரு தனிநிலை தொடரியல் உறுப்பாக நடத்தை செய்யும் மூலமொழியிலுள்ள சொற்கள் குழுமப்பட்டு மூலமொழிக்குப் பொருந்தும்படி அது சொற்களால் பதிலீடு செய்யப்படும் பீன்னுருபை நிர்ணயிக்க சில் விதிகள் உருவாக்கப்பட்டன. சொற்களின் நிரல் வருகை பெரும்பாலும் தக்க வைக்கப்பட்டு சில சமயங்களில் புதிய இணைப்பான்களின் பயன்பாட்டால் நிறைவு செய்யப்படும். சில சமயங்களில் இம்முறை வெளியீட்டில் தொடரியல் தவறுகளுக்கு வழிவக்கும்; இது பற்றி இதை உருவாக்கியவர்கள் கவலைப்படவில்லை. இருப்பினும் குறைந்த அளவு பொருள் பெரும்பாலும் வெளிப்படுத்தப்படும். மேலும் அனுசாரக்காவின் பாகங்கள் அல்லது உறுப்புத் தொகுதிகள் எந்த இயந்திர மொழிபெயர்ப்பிற்கும் அல்லது இயந்திர உதவியுடன் செய்யப்படும் மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கும் பகுதியாக அமையும்.

இவ்வாறு அனுசாரகாவின் வெளியீடு இடைப்பட்ட ஒரு நிலைக்கு முக்கிய வெளியீடாகும் மற்றும் அனுசாரகா போன்ற வெளியீடு ஒரு இரண்டாம் நிலை

விளைவாகும் (by product). இது சீரற்ற மொழிபெயர்ப்பிற்கு கொண்டுசெல்லாததால் மொழிகளின் இணையின் இயந்திர மொழிபெயர்ப்பிற்கும் பயனுள்ள உபாயமாகும்.

இதை உருவாக்கியவர்கள் இதை ஒரு மொழி அணுகி(language acessor) என்று அழைகின்றனர்; இதிலிருந்து முழு நன்மை அடைய பயன்படுத்துபவர்களிடம் சில் பயிற்சியை எதிர்பார்க்கின்றனர். தொடக்கத்தில் கன்னடா, இந்தி மொழிகளுக்கு ஒரு அனுசாரக் ஒழுங்குமுறை அமைக்கப்பட்டு காட்டப்படது. இதன் பின்னர் தெலுங்கு-இந்தி, பஞ்சாபி-இந்தி, மராட்டி-இந்தி, சமஸ்கிருதம்-இந்தி, தமிழ்-இந்தி ஆகிய மொழி இணைகளுக்கு அனுசாரக் ஒழுங்குமுறைகள் உருவாக்கப்பட்டு பார்வையாளர்களுக்கு எடுத்துக்காட்டப்பட்டது.





அனுசாரகா வெளியீட்டில் \

பல மட்டங்கள்

2.7.3 ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு இயந்திர உதவியுடன் மொழிபெயர்ப்பு (ஆங்கிலபாரதி)

ஐ.ஐ.டி.கான்பூரில் மேற்கொள்ளப்பட்ட ஆங்கில பாரதி (*ANGLABHARATI*) திட்டத்தை நிறுவுதில் நான்கு முக்கிய கருதல்கள் இருந்தன. முதலாவது அன்றைய காலகட்டத்தின் நிலையில் சீரான இயந்திர மொழி பெயர்ப்பு சாத்தியமில்லை. எனவே இயந்திர உதவியுடன் மொழிப்பெயர்ப்பு முயற்சிக்கப்பட வேண்டும்; இங்கு இயந்திரம் பெரும்பாலான செயல்களைச் செய்கின்றது. கடினமான வேலைககள் முந்தைய மற்றும் பிந்தைய திருத்தங்களால் (Pre and Post editings) கையாளப்படுகின்றது. இதன்படி ஒவ்வொரு மூலமொழி வாக்கியத்ததையும் தனியாகப் பரிசோதிப்பது மற்றும் முற்கட்டு மற்றும் (anaphora)

வாக்கியங்களுக்கிடையிலான குறிப்புகள் (Intersentential References) போன்ற சிக்கல்களை மனிதத் திருத்துனருக்கு விட்டுவிடுவது என்பன எளிமைக்கு வழி வகுக்கிறது. இரண்டாவது இயந்திர மொழி பெயர்ப்பு, பயனப்படுத்துவோர் மற்றும் பயன்பாட்டு அடிப்படையில் இருக்கவேண்டும் என்ற உணர்வு, ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு உள்ள மொழிபெயர்ப்பை மேற்கொள்ளுவது இந்திய மொழிகளுக்கு இடையிலான இயந்திர மொழிபெயர்ப்பை மேற்கொள்வதைக் காட்டிலும் சிறந்தது என்ற தீர்மானத்திற்கு வழிவகுத்தது. மூன்றாவது ஒரு குறிப்பிட்ட பொருண்மைக்களத்தை அடிப்படையாகக் கொண்ட ஒழுங்குமுறைகள் (domain specitic systems) திறந்த ஒழுங்குமுறைகளைக் காட்டிலும் (open ended systems) உருவாக்குவதற்கு எளியது மற்றும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பு பயன்படுத்துவோர்க்கு சிக்கனமானதாகவும் இருக்கிறது. இறுதியாக வேறுப்பட்ட மொழிகள் ஒற்றுமையுள்ள மொழிகளைக் காட்டிலும் ஆய்வு அடிப்படையில் கூடுதல் அறை கூவல்களை எதிர்கொள்ளும்.

சில முக்கியத்திட்டக் கருத்தல்கள் 90 விழுக்காடு வேலையை இயந்திரத்தாலும் வேலையை மனித முன் திருத்தத்தாலும் (post editing) மேற்கொள்ளும் படிக்கு மொழிப் பெயர்ப்பிற்கு ஒரு பயன்பாட்டு கருவியைத் தருவது என்ற நோக்கம் அடிப்படையில் அமைந்தது, இந்த அமைப்பொழுங்கு படிப்படியாகச் சிக்கலான சூழல்களைக் கையாளும்படிக்கு வளர இயலும், பொத்துமான உரை உருவாக்கும் பகுதிகளின் (text generator modules) இணைப்பால் ஆங்கிலத்திலிருந்து பெரும்பாலான இந்திய மொழிகளுக்கு மொழிப்பெயர்ப்பு ஒரு சீரான இயங்கு முறையைக் கொண்டிருந்தது. மேலும் அதன் பயன்பாட்டிற்கும் விரிவாகத்திற்கும் வசதி செய்ய மனிதனால் இயக்கப்பட்ட மனிதன் - இயந்திர இடைமுகம் கொண்டிருந்தது.

இது ஒரு குழும இந்திய மொழிகளுக்கு (target languages) பயன்படுமாறு போலி இலக்குமொழியை உருவாக்கும் (Pseudo target language) சூழல் வரையறையற்ற இலக்கணம் போன்ற அமைப்புடன் கூடிய அமைப்பொழுங்கை

இலக்காகக் கொண்ட விதி அடிப்படையிலான ஒழுங்குமுறையாகும். பெருந்தரவு ஆய்வின் மூலம் கிடைக்கப்பெற்ற ஒரு குழும விதிகள் போலி மூலமொழிக்கு நகர்வு விதிகளைப் பயன்படுத்தும் சாத்தியமான உறுப்புகளைக் கண்டு கொள்வதற்குப் பயன்படுத்தப்படுகிறது. போலி இலக்குமொழியைப் பயன்படுத்தும் கருத்து, இடைமொழி அணுகுமுறையைப் பயன்படுத்துவதற்குச் சமமான நன்மையைப் பெற அமைப்புசார் ஒற்றுமையை முதன்மையாகப் பயன்படுத்துவதாகும். போலி இலக்குமொழி, இடைமொழி அணுகுமுறையில் பயன்படுத்தப்படும் இடைப்பட்ட மொழி அல்ல இங்கு பொருண்மை உருப்படுத்தத்தை உருவாக்க எந்த முயற்சியும் மேற்கொள்ளப்படவில்லை.

இந்திய மொழிகள் சுதந்திர சொல்குழும நிரலைக் கொண்டு வினையில் முடியும் பண்புடையனவாகும். அமைப்பு ஒற்றுமையின் வழி இந்திய மொழிகளை அவற்றின் மூலம் அடிப்படையில் நான்கு பெரிய குழுமங்களாக வகைப்படுத்தலாம்:

1. இந்திய – ஆரிய மொழிக்குடும்பம் (Indo-Aryan family) (இந்தி, வங்காளம் அஸ்ஸாமிஸ், பஞ்சாபி, மராத்தி, ஒரியா குஜாத்தி போன்றன)
2. திராவிட மொழிக் குடும்பம் (Dravidian Family) (தமிழ், தெலுங்கு, கன்னடம், மலையாளம் போன்ற)
3. ஆஸ்ட்ரோ – ஆசிய மொழிக் குடும்பம் (Astro-Asian family),
4. திபெத்திய பர்மிய மொழிக் குடும்பம் (Tibato-Burman Family).

ஒவ்வொரு குழுமங்களுக்குள்ளும் மொழிகள் உயர்ந்த அளவு அமைப்புப் பொருத்தத்தைக் காட்டுக்கின்றன. ஆங்கில பாரதி திட்டம் இந்த ஒற்றுமையைக் கூடுதல் அளவு இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கப்பயன்படுத்துகிறது.

இலக்குமொழியில் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கு வேண்டி பல பொருண்மை அடையாளங்கள் பயன்படுத்தப்படுகின்றன. தீர்க்கப்படாத பொருண்மை மயக்கங்களின் மாற்றுப் பொருள்கள் போலி இலக்குமொழியில் தக்க

வைக்கப்படுகிறது. ஒவ்வொரு இலக்குமொழிக்கும் உரை உருவாக்கப் பகுதி போலி இலக்குமொழியை இலக்குமொழிக்கு மாற்றுகிறது. ஓவ்வொரு இலக்குமொழிக்கும் திருத்துவான் (corrector) பயன்படுத்தப்படுகிறது. இறுதியாக மனித இயக்கத்தால் தூண்டப்பட்டபின் திருத்தும் தொகுதி (Post editing package) இறுதி திருத்தங்களைச் செய்வத்றகுப் பயன்படுத்தப்படுகிறது. பின்திருத்தி இலக்குமொழியை மட்டும் அறிந்தால் போதுமானது.

ஆங்கில பாரதி திட்டம் சின்ஹா (Sinha, 1993) என்பவரால் 1991-ல் உருவாக்கப்பட்டது. இத்திட்டத்தில் ஆங்கிலத்திலிருந்து இந்தி மற்றும் தெலுங்கு மொழிகளுக்குச் செயல்படக்கூடிய மூலமுன்மாதிரி மொழிபெயர்ப்பு (Functional Prototype) ஒழுங்குமுறை உருவாக்கப்பட்டது. பின்வரும் திட்டவரைபடம் இந்த மொழி பெயர்ப்பு ஒழுங்குமுறையை விளக்கும். பின்வருவன ஆங்கில பாரதி மொழிபெயர்ப்புத் திட்டத்தின் முக்கியக்கூறுகளாகும்.

1. விதி அடிப்படை

விதி அடிப்படை (Rule base) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு வாக்கியங்களின் அமைப்புகளைப் பொருத்துவதற்கான விதிகளைக் கொண்டிருக்கின்றது. இந்த ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு அமைப்பு ஒழுங்கு மாற்றங்களின் தரவு மையம் மொழிபெயர்களுக்கு அமைப்பு ஒழுங்கு மாற்றங்கள் தரவு மையம் மொழிபெயர்க்கவேண்டிய வாக்கியத்திற்குகாகக் கிளை அமைப்பைப் பெறும் வேலையைப் புறக்கணித்து புறக்கிளைகளிலிருந்து புறக்கிளை மாற்றத்தை செய்யும் வேலைக்குப் பணிக்கப்படுகிறது. வாக்கியங்கள் பொருண்மையைக் கண்டு பிடிக்க புற அமைப்பொழுங்கைப் பயன்படுத்தும் கருத்து, மொழியியலில் மிகப் பழமையானதாகும். இந்த அணுகுமுறை எளிதானது என்றாலும் இத்திட்டம் ஒரு மொழியின் புற அமைப்பு ஒழுங்களின் தனிப்பட்ட தன்மைகளைக் கண்டுபிடிக்கச் செயலூக்கமுடையதாகும். ஆங்கில பாரதியில் பயன்படுத்தப்படும் அணுகுமுறை சாம்ஸகியின் தொடரமைப்பு

இலக்கணத்திலிருந்தும் சொல்சார்- செயல்பாடு சார் இலக்கணத்தின் (Lexical functional Grammer) அமைப்பிலிருந்தும் (C-structure) அதிகமாக எடுத்தாண்டுள்ளது. விதி அடிப்படை (rule base) என்று கூறப்படும் ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கான அமைப்பு மாற்ற விதிகளின் தரவு மூலம் ஆங்கில பாரதி ஒழுங்குமுறையில் மையப் பகுதியாக அமைகிறது. இது ஆங்கிலத்திலிருந்து மொழிபெயர்கும் போது தொடரியலில் ஏற்படும் முக்கிய மாற்றங்களைக் கவனித்துக்கொள்கிறது. முன்னர் கூறியப்படி இந்திய மொழிகளுக்குப் பொது விதி அடிப்படையை உருவாக்கி ஆங்கில பாரதி ஆங்கிலத்திலிருந்து மொழிபெயர்க்கும் போது சிறந்த நன்மையை வெளிப்படுத்துகிறது.

2. அர்த்த மயக்க நீக்கி

அர்த்த மயக்கநீக்கி (Sense disambiguator) மூல மொழியின் ஒவ்வொரு சொல்லுக்கும் சரியான அர்த்ததை தேர்ந்தெடுக்கப் பொறுப்புள்ளதாகும். இங்கு அர்த்த மயக்கங்க நிக்கம் மூல மொழியின் உரையில் மட்டுமே நிகழ்த்தப் பெறுகின்றது. ஆங்கில பாரதியில் பயன்படுத்தப்படும் இவ்வணுகுமுறையை விதியால் விதி-பொருகோள் (Rule-by-rule Semantic Interpretation) எனக் கூறலாம். ஒரு தொடரியல் விதி பயன்படுத்தப்படும் ஒவ்வொரு நேரத்திலும் பொருள்கோளி (Semantic Intepreter) பயன்படுத்தப்படுகின்றது.

3. இலக்கு உரை உருவாக்கி

இலக்கு உரை உருவாக்கி (Target text generator) மொழிபெயர்ப்பு ஒழுங்குமுறையின் இறுதியாக அமைகின்றது. இதன் செயல்பாடு பொறுத்தமுறும் மூல மொழிகளிலிருந்து மொழிபெயர்க்கப்பட்ட வெளியீட்டை உருவாக்குவது ஆகும். இது ஆங்கில பாரதியின் முந்தைய நிலைகளின் உருவாக்கப்பட்ட இடைப்பட்ட வடிவத்தை உள்ளீடாக ஏற்றுக் கொள்கின்றது. இவ்வேலையை

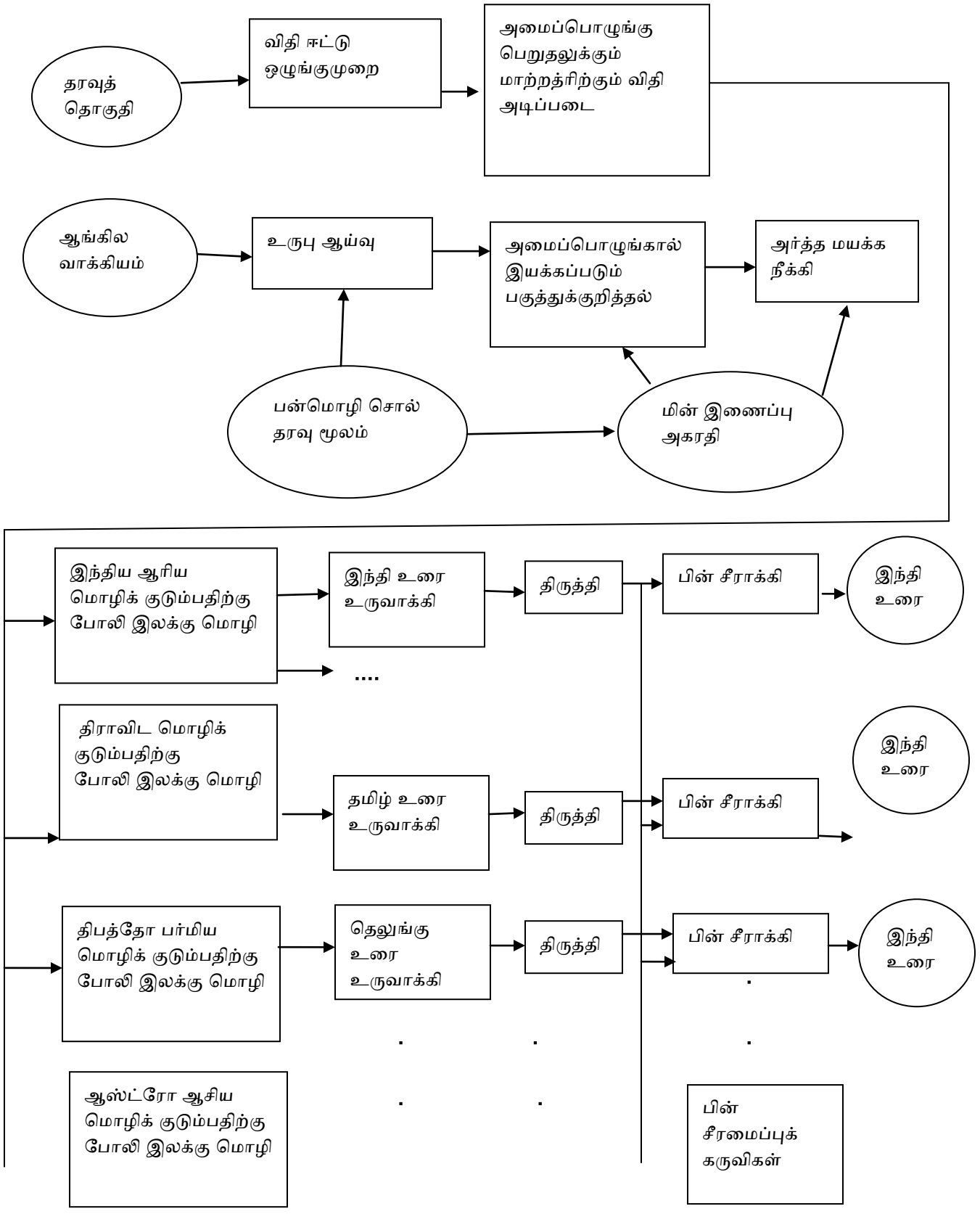
இயற்கை மொழி உருவாக்கம் (Natural Language Generation) என்று அழைக்கப்படும் வேலையிலிருந்து வேறுபட்டவையாகும்; இவ்வர்த்ததில் பிந்தையது எனக் கூறவேண்டும் என்பதுடன் திட்ட நிலை (Strategic level) எவ்வாறு கூற வேண்டும் என்பதையும் நடவடிக்கை நிலை (Tactical Level) தீர்மானிக்க வேண்டும். ஒரு விதி அடிப்படை மற்றும் அர்த்த மயக்க நீக்கி இவற்றைப் பயன்படுத்தும் வேறுபட்ட உரை உருவாக்கிகளைக் கொண்டு பல இலக்கு மொழிகளுக்கும் பொது இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை அமைப்பு கிடைக்கப்பெறுகின்றது.

4. பன்மொழி அகராதி

பன்மொழி அகராதி (Multilingual Dictionary) தொடரியல் வகைப்பாடுகள், சாத்தியமான அர்த்தங்கள், அர்த்தங்களின் மயக்கத்தை நீக்கும் முக்கிய கூறுகள் (Keys) இவற்றை உள்ளடக்கிய ஆங்கிலச் சொற்களுக்கும் இணையாக மூல மொழிச் சொற்களுக்குப் பல விளக்கங்களைக் கொண்டிருக்கும்.

5. விதி அடிப்படை ஈட்டி

விதி அடிப்படை ஈட்டி (rule base acquirer) மொழிபெயர்ப்பு ஒழுங்குமுறைக்கு விதி அடிப்படையை உருவாக்குகிறது.


தரவுத் தொகுதி
விதி ஈட்டு ஒழுங்குமுறை
அமைப்பொழுங்கு பெறுதலுக்கும் மாற்றத்திற்கும் விதி அடிப்படை
ஆங்கில வாக்கியம்
உருபு ஆய்வு
அமைப்பொழுங்கால் இயக்கப்படும் பகுத்துக்குறித்தல்
அர்த்த மயக்க நீக்கி
பன்மொழி சொல் தரவு மூலம்
மின் இணைப்பு அகரதி
இந்திய ஆரிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
இந்தி உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
....
திராவிட மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தமிழ் உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
திபத்தோ பர்மிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
தெலுங்கு உரை உருவாக்கி
திருத்தி
பின் சீராக்கி
இந்தி உரை
ஆஸ்ட்ரோ ஆசிய மொழிக் குடும்பதிற்கு போலி இலக்கு மொழி
பின் சீரமைப்புக் கருவிகள்

2.8 தமிழ் சார் இயந்திர மொழிப்பெயர்ப்பு

தமிழ்சார் இயந்திர மொழிபெயர்ப்புகள் பல மேற்கொள்ளப்பட்டு தமிழி நாட்டில் இயந்திர மொழிபெயர்ப்பின் வளர்ச்சிக்கு வித்திட்டன.

2.8.1. தமிழ் சார் இயந்திர மொழிபெயப்பு முயற்சிகள்

2.8.1.1 தமிழ் உருஷ்யன் மொழிப் பெயர்ப்புத்திட்டம்

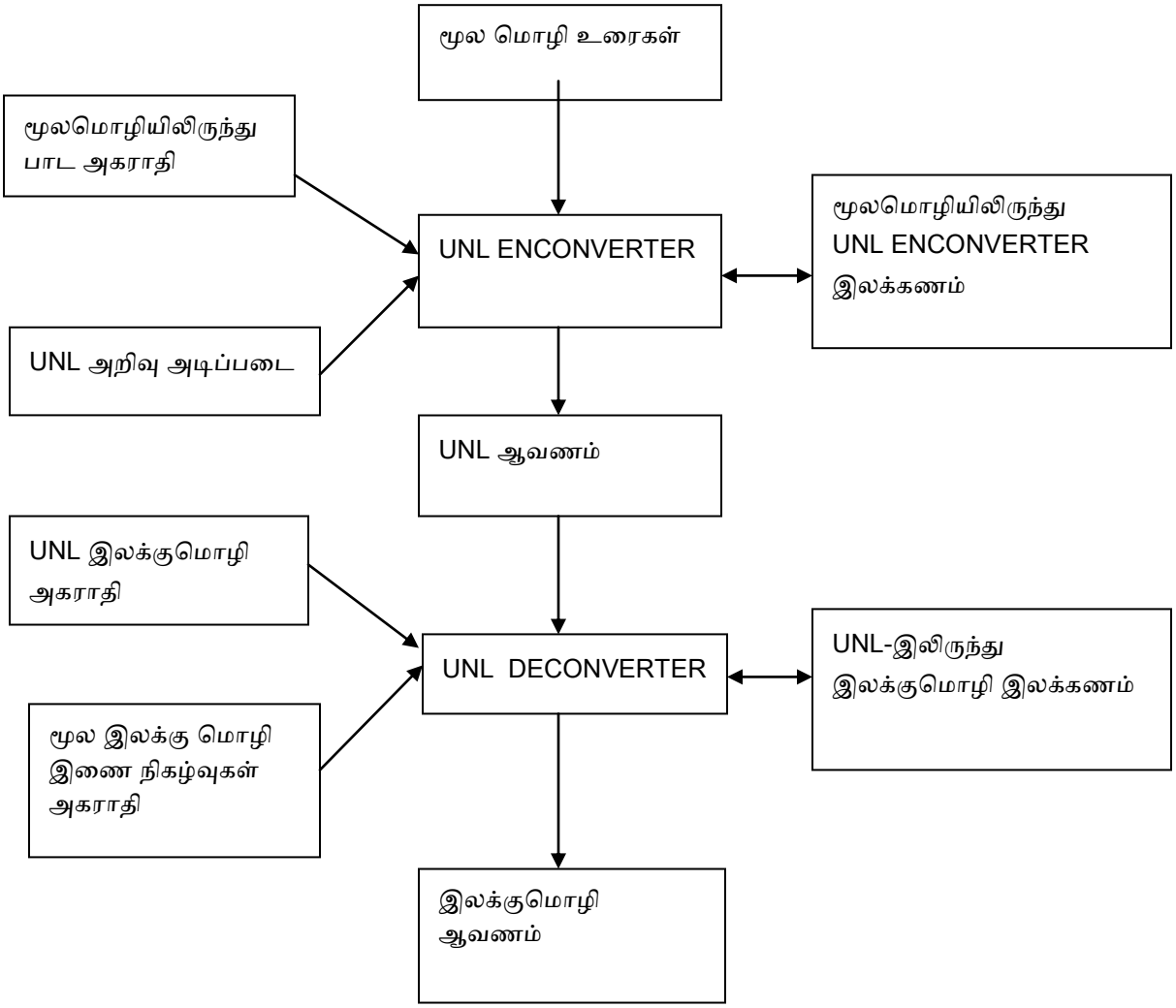
1985-ஆம் தமிழ் பல்கலைக்கழகத்தில் (Tumtst Tamil University Machine Inaslation System) என்ற திட்டப்பணி உருவாக்கப்பட்டது. இதன் மூலம் ருஷ்ய மொழியில் உள்ள செய்திகளைத் தமிழில் மொழிப்பெயர்க்க ஒரு மென்பொருள் உருவாக்கப்பட்டது. இத்திட்டம் தமிழ்ப்பல்கலைக்கழகம் கண்ணி அறிவியல்துறை, மொழியியல் துறை மற்றும் மொழிப் பொயர்ப்புத் துறை ஆகியவைகளின் கூட்டு முயற்சியில் உருவானதாகும்.

2.8.1.2 உலக வலைப்பின்னல் மொழி – தமிழ்க்கான இடைமொழி இயந்திர மொழிப்பெயர்ப்பு

இந்த அணுகுமுறை சென்னையில் உள்ள அண்ணா பல்கலைக்கழகத்தில் கணிப்பொறி மொழியியல் துறையில் மேற்கொள்ளப்பட்டது. (Dhanabalan & Geetha, 2004). இந்த அணுகு முறையில் இடையீட்டு மொழியின் உருப்படுத்தம் பயன்படுத்தப்படுகிறது. மூலமொழி அகராதியையும் இலக்கணத் தகவல்களையும் பயன்படுத்தி இடைமொழி உருப்படுத்தம் மூலமொழியிலிருந்து இலக்கு மொழிக்கு மொழிப்பெயர்ப்பு செய்ய பயன்படுத்தப்படும் இடைமொழி அமைப்பு உலகளாவிய வலைப்பின்னல் மொழி (Universal Networking Language (UNL)) ஆகும். இம்மொழி இடைமொழி அமைப்பாகத் திட்டமிடப்பட்டு மூலமொழியிலிருந்து

இலக்கு மொழிக்கும் இலக்கு மொழியிலிருந்து மூலமொழிக்கும் தானியங்கி இயந்திர மொழிபெயர்ப்புயைச் சாத்தியமாக்குகின்றது. UNL-ஐ பயன்படுத்திய மொழிபெயர்ப்பு ஒழுங்குமுறை மூலமொழியிலிருந்து UNL-க்கு மாற Encoverter-ஐயும் UNL-இலிருந்து இலக்கு மொழிக்கு மாற Deconverter பயன்படுத்துகின்றது.

இந்த அணுகுமுறையின் திட்டவடிவம் கீழே தரப்பட்டுள்ளது.



மேற்கண்டவாறு படத்தில் காட்டியவாறு UNL அமைப்பு வேலை செய்து UNL மொழியிலிருந்து தமிழுக்கு மொழிபெயர்ப்பு செய்யும்.

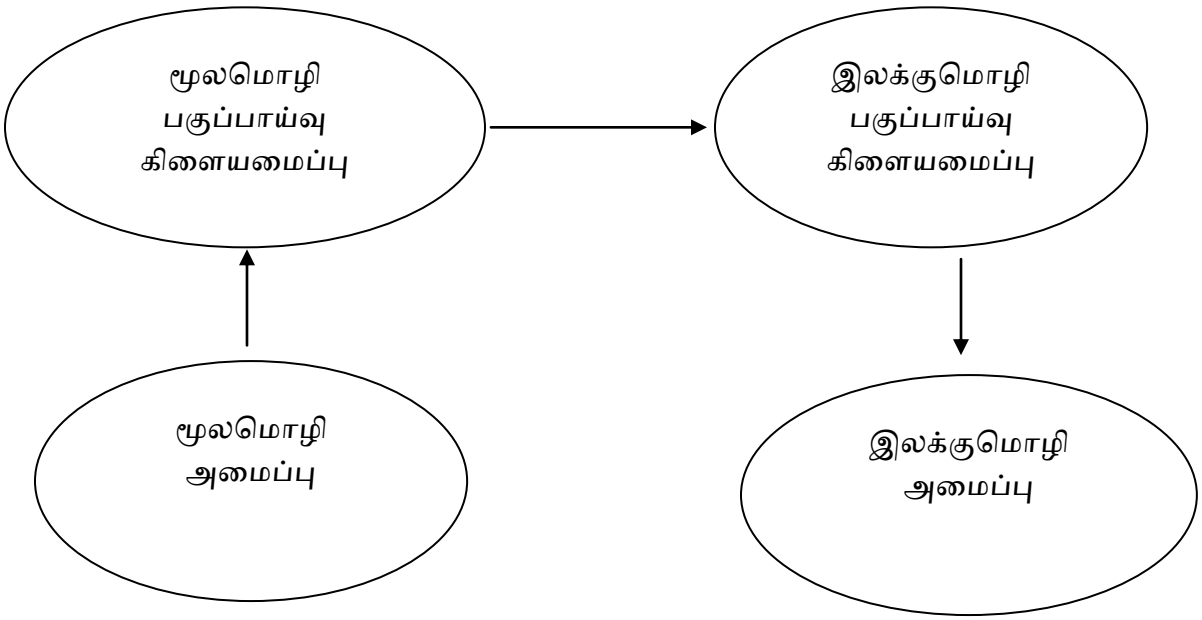
2.8.1.3 ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்கும் திட்டம்

ஆங்கிலத்திலிருந்து மொழியியல் புத்தகங்களைத் தமிழில் மொழிபெயர்க்க ஒரு இயந்திர மொழி பெயர்ப்புக் கருவி உருவாக்க, காமாட்சி மற்றும் ராஜேந்திரன் (Kamachi Rajendran, 2004) ஒரு திட்டத்தை உருவாக்கி செயல்படுத்த முயன்றனர்.

தமிழில் முதுகலை மொழியியல் பாடத்திட்டத்திற்கு வேண்டிய மொழியியல் புத்தகங்களைத் தமிழில் மொழிப்பெயர்க்க வேண்டிய கட்டாயம் ஏற்பட்டதால் இத்தகைய ஆய்வு மேற்கொள்ளப்பட்டது. இவ்வாய்வு மொழியியல் புத்தகங்களை ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்க்க உதவும் ஒரு கருவி தயாரிப்பதை நோக்கமாக கொண்டது. இக்கருவி பின்வரும் நான்கு பகுதிகளைக் கொண்டது.

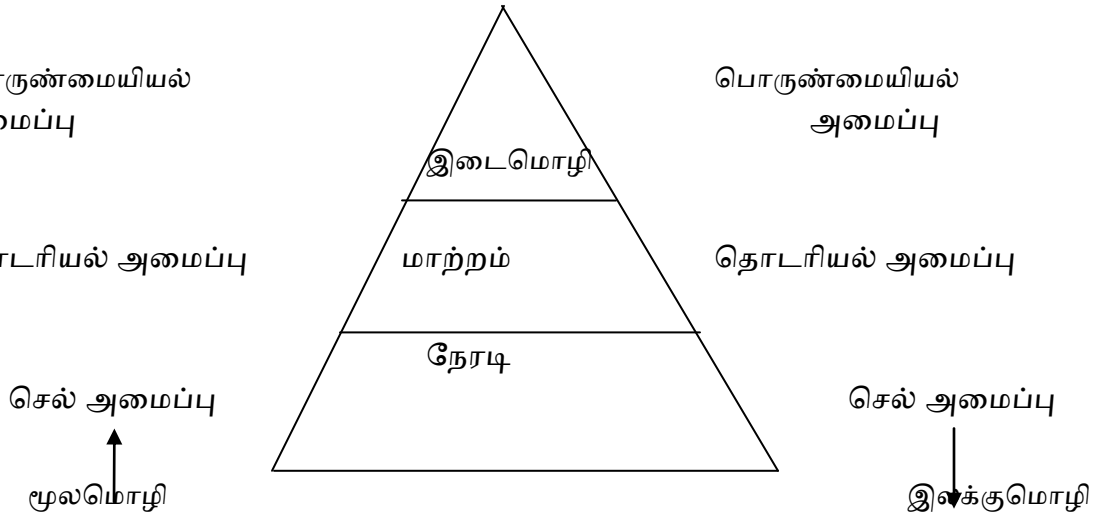
1. தமிழ் மொழி அமைப்புப் பகுத்தாய்வு
2. ஆங்கில மொழி அமைப்புப் பகுத்தாய்வு
3. ஆங்கில - தமிழ் மொழிபெயர்ப்பு அகராதி
4. ஆங்கில – தமிழ்மொழி மாற்றமைப்பு விதிகள்
5. ஆங்கில மொழியமைப்பு உருவாக்கி

இந்த மொழிப்பெயர்ப்பு கருவியின் செயல்பாட்டை விளக்கும்



ஒரு ஒழுங்குமுறையை உருவாக்குவதில் ஏற்படும் சிக்கலின் நிலை மொழி பெயர்க்கப்படும் இரு மொழிகளின் ஒற்றுமை மற்றும் வேற்றுமைகளை பொறுத்து

அமையும். இவ்வொற்றுமை வேற்றுமைகள் சொல்நிலையில் மொழி மாற்றம் செய்யப்பட வேண்டுமா? தொடர் நிலையில் மாற்றம் செய்யப்பட வேண்டுமா? பொருண்மை நிலையில் மாற்றம் செய்யப்பட வேண்டுமா என்பதைத் தீர்மானிக்கும் பின்வரும் படம் இச்சாத்தியங்களை வெளிப்படுத்தும்.



தமிழிலிருந்து பிற மொழிகளுக்கு மற்றும் பிற மொழிகளிலிருந்து தமிழ்மொழிக்கு இயந்திர உதவி மொழிபெயர்ப்பு ஒழுங்குமுறைகள் தமிழ்மொழியின் மீது பல நிலைகளிலான ஆய்வுகளை உள்ளடக்கும். ஒவ்வொரு நிலைக்கும் அதற்கான சிக்கல்கள் காணப்படும்.

2.8.2 தமிழ்சார் மொழிபெயர்ப்பு நடவடிக்கைகள்

2.8.2.1 உருபனியல் ஆய்வு (Morphological Analysis)

இது வேர் சொற்களைக் கண்டுபிடிப்பதையும் அதன் இலக்கண பண்புப் கூறுகளைக் கண்டுபிடிப்பதையும் உள்ளடக்கும். தமிழ் ஒரு திரிபு (Inflection) வளமுள்ள மொழி ஆகையால் அதன் பகுப்பாய்வு ஆங்கிலத்தின் உருபனியல் பகுப்பாய்வை விடச் சிக்கல் வாய்ந்ததாக அமையும். புணர்ச்சி விதிகள் தமிழில் உருபனியல் ஆய்வி உருவாக்குவதை கடினமாக்குகின்றது. சொற்களை உருபன்களாகப் பிரிப்பதிலும் அவற்றிற்குப் பொறுத்தமான பொருண்மை மற்றும்

இலக்கணத் தகவல்கள் தருவதிலும் கவனம் தேவை. பின்வரும் எடுத்துக்காட்டு இதைக் தெளப்வுப்படுத்தும்.

தமிழ்சொல் : பணத்திற்காகத்தானே

உருபனியல் பிரிப்பு : பணம் + க்கு + ஆக + தான் + ஏ

வேர் சொல் : (பணம்)

நான்காம் வேற்றுமை உருபு (க்கு)

கொடை வேற்றுமை உருபு (ஆக)

இச்சொல் ஆங்கிலத்தில் மொழிப்பெயர்க்கப்படும் போது இது ஒரே சொல்லாக மொழி பெயர்க்கப்படாமல் கடிச வாந ளயமந அடிஇநல டிஇடல என மொழி மொழிபெயர்க்கப்பட வேண்டும்.

2.8.2.2 சொல் வகைப்பாட்டு அடையாளப்படுத்தல் (Parts of Speech Tagging)

சொல் வகைப்பாட்டு அடையாளப்படுத்தி வாக்கியத்திற்கும், சொற்களுக்கும் சொல்வகைப்பாடு தரவேண்டும். இது வாக்கியத்தின் அமைப்பைப் புரிந்துக்கொள்ளவும் வாக்கியத்தின் அமைப்பை உருவாக்கவும் முக்கியமானதாகும்.

தமிழ் வாக்கியம் : அவன் படி ஏறினான்

He staircase Climbed

மேற்கண்ட வாக்கியம் பின்வருமாறு அடையாளப்படுத்தப்படும்.

அவன் < PRO >

படி < N+PRO >

ஏறினான் < V+PST+3SM >

இதில்,

PRO ⇒ Pronoun ‘மாற்றுப் பெயர் ’

N ⇒ Noun ‘ பெயர் ’

NOM ⇒ Nominative ‘ எழுவாய்ப் பெயர் ’

V ⇒ Verb ‘ வினை ’

PST ⇒ Past tense ‘ இறந்த காலம் ’

3SM ⇒ 3 Person Singular, Masculine ‘படர்க்கை ஆண்பால் ’

இந்த வேர்ச்சொல்லான ‘படி’ ‘staircase’ என்ற பொருளையோ ‘read’ என்ற பொருளையோ தரலாம். சோல்வாகைப் பாட்டு அடையாளப்படுத்தி சொல் மற்றும் தகவலைப் பயன்படுத்தி இந்த பொருள் மயக்கத்தைத் தீர்க்கும். இவ்வாறு ‘படி’ என்பதற்குப் பெயர் சொல்வகைப் பாட்டு அடையாளம் தரப்படும்.

2.8.2.3 தொடரியல் ஆய்வு (Syntax analysis)

தொடரியல் ஆய்வின் போது பகுத்தாப்படுகையில் வாக்கிய அமைப்பு தெரிந்து கொள்ளப்படும். பகுத்தாய்வதால் எச்சத்தொடர் எல்லைகள், முன்னுருபு இணைப்பு அடைகள் என்பன தெரிந்துக்கொள்ளப்பட்டு அவற்றின் பொருத்தமான உறுப்புகளுடன் தொடர்புப் படுத்தப்படும். தமிழ் ஒரு சொல் சுதந்திரமான மொழியாகும். இது தொடரியலில் பகுத்தாய்தலைச் சிக்கலாகும். பெயருடன் இணைந்து வரும் வேற்றுமை உருபுகள் வாக்கியத்தில் இவ்வுறுப்புகளின் பங்களிப்பைக் கண்டு பிடிக்கின்றன. ஆனால் வேற்றுமை உருபு இல்லாது இருப்பு அமைப்பு சார் பொருள் மயக்கத்திற்கு வழிவகுக்கும். எடுத்துக்காட்டாக,

தமிழ் வாக்கியம் : ‘அவள் அக்காள் வீட்டிற்குச் சென்றாள்’

‘ She / he Sister house – dative go+pst 2sf’

என்பதற்கு இரண்டு பகுத்துக் குறிப்புகள் கிடைக்கும் :

(அவள்) NP (அக்காள் வீட்டிற்கு) NP சென்றாள் V

(அவள் அக்காள்) NP (vii TT : RKU) NP சென்றாள் V

2.8.2.4 பொருண்மையியல் ஆய்வு (Semantic Analysis)

பல அர்த்தங்கள் உள்ள சொற்களுக்கு ஒரு குறிப்பிட்ட சூழலுக்குத் தகுந்தவாறு சரியான அர்த்தம் கண்டுபிடிக்கப்பட வேண்டும். இலக்குமொழி நிகரன் மயக்கம் நீக்கப்பட அர்த்தம் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும்.

தமிழ் வாக்கியம் 1 : அவன் படி ஏறினான்

Climb + pst + 3sm

தமிழ் வாக்கியம் 2 : அவன் படியால் அளந்தான்

He Vessltins Measuvet Ps ++3sm

மேற்கூறிய வாக்கியங்களில் படி என்பதற்கு இரண்டு அர்த்தங்கள் இருக்கின்றன.

1. Staircase
2. A vessel used to measure rice and whet.

முதல் வாக்கியத்தில் ஏறினான் என்ற சொல்லுடன் இணைந்து வருகையைப் பயன்படுத்தி படி என்பதன் அர்த்தம் ‘Staircase’ என்பதுதான் என்றும் இரண்டாம் வாக்கியத்தில் அளந்தான் என்ற சொல்லுடன் சேர்ந்து வருகையால் படி என்பதன் அர்த்தம் ‘vessel’ என்பதுதான் என்றும் நிர்ணயிக்கப்படுகின்றது.

2.8.2.5 தொடரியல் சார் மாற்றம் (Syntactic Transfer)

மொழிபெயர்ப்பு செல்பாட்டில் மாற்றல் அணுகுமுறை (transfer approach) பயன்படுத்தப்படும் போது தொடரியல் மாற்றப்பகுதி தேவைப்படும். ஊள்ளீடு

செய்யப்படும் வாக்கியத்தின் அமைப்பு இலக்குமொழியின் பொருத்தமான அமைப்பிற்குத் தகுந்தவாறு மாற்றப்படுகிறது.

தமிழ் வாக்கியம் : ராமன் புத்தகம் படித்தான்

Raman book read

அமைப்பு: Noun Noun Verb

மேற்கண்ட வாக்கியங்களுக்கு நிகரான வாக்கியங்கள் கீழ்வருவன.

ஆங்கில வாக்கியம் : Raman read a book

அமைப்பு :Noun Verb Noun

இந்த மொழிபெயர்ப்பில் Noun Noun Verb என்ற தமிழ் வாக்கியம் Noun Verb Noun என்று ஆங்கிலத்திற்கு நிகராக மாற்றப்படுகிறது. இந்த எடுத்துக்காட்டு மிக எளிமையானதாகும். ஆனால் இச்செயல்பாடு கலவை வாக்கியங்களுக்கு கிடையில் கடினமானதாகும்.

2 .8.2.6 அகராதி பொருத்தம் (Dictionary Mapping)

இருமொழி அகராதியைப் பயன்படுத்தி மூலமொழி வேர் சொற்கள் இலக்குமொழியில் அதற்கு நிகரான சொற்களுடன் பொருத்தம் செய்யப்படும். ஒரு மூலமொழி வேர்ச்சொல்லுக்கு இலக்கு மொழியில் பல சொற்கள் கொண்ட தொடர்கள் நிகரன்களாக வரலாம். சூழல் அடிப்படையில் பொருத்தமான சொல் சூழல் அடிப்படையில் தேர்ந்தெடுக்கப்பட்ட வேண்டும். வேர்ச் சொல்லின் இலக்கண பண்புக் கூறுகளின் இலக்கணம் அகராதியில் அடங்கி இருக்கும்.

2.8.2.7 உருபனியல் உருவாக்கம் (Morphdogical Generation)

இங்கு சொல்லின் திரிபு வடிவம் வேர் மற்றும் அதன் இலக்கண பண்புக் கூறுகளால் உருவாக்கப்படும்.

தமிழ் வேர்ச்சொல் : 'படி'
இலக்கண பண்புக் கூறுகள் : 'Past tense'+ 3rd person, Singular, Mascuine'

சொல் வடிவு : ' படித்தான் '

2.8.3 தமிழ்மொழி ஆய்வுக்குத் தேவையான மூலவளமும் கருவிகளும்

2.8.3.1 கருவிகள் (Tools)

2.8.3.1.1 உருபனியல் ஆய்வி (Morphoogical Analyzer)

95 விழுக்காட்டிற்கு மேல் முழுமையுடன் தொழில் நுட்பத்தின் தற்போதைய நிலையைப் பயன்படுத்தி ஒரு திறமையான உருபனியல் ஆய்வி தேவை. தற்போதைய உருபனியல் ஆய்விகள் மேற்கோள்வாய்வாய்பாட்டு (Paradim approach) அணுகுமுறையையும் முற்றுநிலைத் தானியங்கியையும் (Finite State Automata) பயன்படுத்துகின்றனர்.

2.8.3.1.2 அடையாளப்படுத்தி

வாக்கியங்களை அடையாளப்படுத்த முழுதானியக்க, பகுதிதானியக்க, விதி அடிப்படையிலான அல்லது புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தித் தேவை. இவ்வகையான அடையாளப்படுத்தல் இயந்திரமொழி பெயர்ப்புக்கு மட்டுமல்லாமல் தகவல் பிரித்தெடுப்பு, தகவல் மீட்பு உரை சுருக்கம் மற்றும் பிற பல்வேறுபட்ட பயன்பாடுகளிலும் பயன்படுகிறது.

2.8.3.1.3 பகுத்துக் குறிப்பான் (Parser)

ஒரு பகுத்துக் குறிப்பான் சார்பு இலக்கணம் (dependency Gramman) தொடரமைப்பு இலக்கணம் (Phrase Structure Grammer), கிளை இணைப்பு இலக்கணம் (Tree Adjoining Grammar) போன்ற ஏதாவது இலக்கண வடிவமைப்புகள் அடிப்படையில் வாக்கியங்களுக்குப் பகுப்புக் கிளை தரும். ஒரு பகுத்துக் குறிப்பான் தமிழ்மொழிப் பகுப்பாய்விற்கு மிகத் தேவை. இது விதிகள் மற்றும் பகுத்துக் குறிக்கும் இயந்திரத்தை உருவாக்குதலையும் உள்ளடக்கும். சில இயற்கை மொழி ஆய்வுப் பயன்பாடுகள் முழு பகுத்துக்குறித்தலை வேண்டாது, அவைகளுக்குப் பகுதி பகுத்துக் குறித்தல் போதுமானது. பெயர்த்தொடர் கண்டுபிடிப்பு, எச்சத்தொடர் கண்டுபிடிப்பு என்பன இந்தப் பயன்பாடுகளுக்கு உதவும்.

2.8.3.1.5 பொருண்மையியல் ஆய்வி (Semantic Analyser)

சொல் அர்த்த மயக்கம் நீக்கும் கருவியை உருவாக்கும் வேலை அதன் தொடக்க நிலையில் தான் இருக்கிறது. இந்த வேலை கூடுதலான கவனக் குவிப்பையும் கடினமான முயற்சிகளையும் வேண்டும். மேலும் இது சொற்கள் இணைந்து வரும் தகவலைச் சேகரிக்க வேண்டி பெரிய அடையாளப்படுத்தப்பட்ட விரிதரவை வேண்டும்.

2.8.3.1.6 உருபனியல் உருவாக்கி (Morphological Generator)

நம்மிடம் உருபனியல் ஆய்வுக்கு நன்றாக அமைக்கப்பட்ட முற்றுநிலை இயந்திரம் இருந்தால் அதை உருபனியல் உருவாக்குதலுக்குச் சிறிது மாற்றமின்ற அல்லது சிறிது மாற்றத்துடன் அதே இயந்திரத்தைப் பயன்படுத்தலாம். இவ்வாறு

உருவாக்கப்பட்ட உருபனியல் உருவாக்கிகள் திருப்தியான விளைவுகளைக் காட்டுகின்றன.

2.8.3.2 மூலவளங்கள் அகராதி (Resources)

2.8.3.2.1 அகராதி (Dictionary)

பல அகராதிகள் தமிழில் இருந்தாலும் அவையெல்லாம் மின்வடிவில் இல்லை. இந்த அகராதிகள் இயந்திரம் படிக்கவியலும் வடிவத்தில் இருந்தால்தான் அவற்றைக் கணினி செயல்பாட்டிற்குப் பயன்படுத்த இயலும்.

2.8.3.2.2 மொழி கடந்த அகராதி (Cross -lingual Dictionary)

இது சாதாரண மொழிகள் கடந்த அகராதியைக் குறிப்பிடவில்லை. இயந்திர மொழிபெயர்ப்புக்கு நமக்கு தேவையான அகராதி மொழி கடந்த வேர் அகராதிகளாகும். தனிப்பட்ட மனிதர்களும் நிறுவனங்களும் அவர்களுடைய பயன்பாட்டிற்காக இருமொழி அகராதிகளை உருவாக்குகின்றனர். இவை இந்த ஆய்வு களத்தில் சூடுபடும் ஆய்வாரள்களுக்கிடையில் பங்கிடப்படவேண்டும்.

2.8.3.2.3 தரவுத்தொகுதி (Corpus)

ஒரு மொழி தரவுத்தொகுதி அம்மொழியின் நடை, சொல், பயன்பாடு போன்றவற்றின் உருப்படுத்தம் ஆகும். இது மேலும் உருபனியல் பகுப்பாவிகள், பகுத்துக் குறிப்பான்கள் மற்றும் உருவாக்கிகள் என்பனவற்றை மதிப்பீடு செய்யும் பரிசோதனைக் கருவியாகச் செயல்படும். இந்திய மொழிகளின் மைய நிறுவனம் சுமார் 15 மொழிகளுக்கு மூன்று மில்லியன் சொற்களைக் கொண்ட தரவுத்தொகுதிகளை உருவாக்கியுள்ளது. மேலும் சில தரவுத்தொகுதிகளை அடையாளப்படுத்தி உள்ளது. இருக்கின்ற வசதிகளை வைத்துக்கொண்டு

பல்வேறுபட்ட மொழிகளுக்கு விரிதரவைச் சேகரிக்க வேண்டி பொதுவான திட்டம் ஊக்கப்படுத்தப்பட வேண்டும். புத்தகங்கள், நாளிதழ்கள், கால இதழ்கள் போன்றவற்றஇ தவிர இணையதளங்களைப் பயன்படுத்தி தரவுத்தொகுதி சேகரிக்கப்பட்ட வேண்டும்.

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி

சொல்வகைப்பாடு அடையாளப்படுத்தப்பட்ட விரிதரவில் (POS Tagged Corpus) உள்ள எல்லா வாக்கியங்களும் சொல்வகைப்பாட்டிற்கு வேண்டி அடையாளப்படுத்தப்பட்டிருக்க வேண்டும். இந்த அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி, புள்ளியியல் அடிப்படையிலான சொல் வகைப்பாட்டு அடையாளப்படுத்தி உருவாக்குவதற்குத் துணைபுரியும். இம்மூலவளத்தின் மீது பல்வேறு வகையான ஆய்வுகளினால் ஒரு மொழிக்குக் கணினி சார் இலக்கணம் உருவாக்க இயலும்.

இணை தரவுத்தொகுதிகள் (Parellel Corpora)

இணைதரவுத்தொகுதி இருமொழிகளுக்கு மாற்றமைவு விதிகளை (transfer rules) உருவாக்க முக்கியமான மூலவளமாகும். இது புள்ளியியல் அல்லது எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறைகள் உருவாக்கத் துணைபுரியும். இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளை உருவாக்குவதற்கு முன் வரிசையாக்க விதிகள் இணைவிரிதரவிகளின் பத்திகள், வாக்கியங்கள் சொற்கள் இவற்றை வரிசையாக்கம் செய்ய வரிசையாக்கக் கருவிகள் தேவைகள். சுப்ல வெளிப்யீட்டாளர்களிடமும் அரசு நிறுவனங்களிலும் இம்மாதிரியான இணைதரவுத்தொகுதிகள் இருக்கின்றன. இந்த இணைதரவுத்தொகுதிகள் விளைவாக்கம் உள்ள ஒழுங்குமுறைகளை உருவாக்க வேண்டி ஆராய்ச்சியாளர்களுக்குப் பயன்படுத்தத் தரப்படவேண்டும்.

2.8.3.2.4 மாற்றமைவு இலக்கணம்

மாற்றமைவு இலக்கணம் (transfer grammar)இருமொழிகளின் தொடரியல் அமைப்புக்கு இடையில் உள்ள முரண்வேறுபாடுகளை அறிந்து கொள்கின்றது. மாற்றமைவு அணுகுமுறைகளை உருவாக்க இம்மாதிரியான மாற்றமைவுப் பகுதியை உருவாக்குவது கட்டாயமாகும். மொழிகளைப் பற்றிய ஆழ்ந்த அறிவுள்ள மொழியியலார் இம்மாதிரியான மூல வளத்தை உருவாக்கலாம்.

2.8.3.2.5 இணைவமைதி அகராதி

இணைவமைதி அகராதி (Collocation dictionary) மொழியின் மயக்கமான சொற்களின் முறையான சேர்க்கைகளைக் கொண்டிருக்கும். அம்மாதிரியான அகராதி அர்த்த மயக்க நீக்கிற்கு முக்கியமான பகுதியாகும். தமிழுக்கு அம்மாதிரியான மூலவளத்தைச் சில ஆய்வாளர் குழுமங்கள் செய்து வருகின்றன.

2.9 நிறைவுரை

இவ்வியலில் கணிப்பொறி வழி இயந்திர மொழிபெயர்ப்பின் தேவை, கணீபொறி வழி மொழிபெயர்ப்பின் வகைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் சிக்கல்கள், கணிப்பொறி வழி மொழிபெயர்ப்பின் எல்லைகள், கணிப்பொறி வழி மொழிபெயர்ப்பின் நிறைகளும் குறைகளும், இந்தியாவில் இயந்திர மொழி பெயர்ப்பின் வளர்ச்சி, தமிழ்ச்சார்ந்த இயந்திர மொழிபெயர்ப்புகள் என்ற தலைப்புகளில் விளக்கங்கள் தரப்பட்டுள்ளன.

இயல் 4
தமிழ்த் தரவுத்தொகுதியைச் சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தல்

4.1 அறிமுகம்

தரவுத்தொகுதிகளைச் சொல்வகைப்பாட்டுக்கு அடையாளப்படுத்துதல் முக்கியமான செயல்பாடாகும். சொல்வகைப்பாடு அடையாளப் படுத்தப்படாத தரவுத்தொகுதிகள் இயற்கை மொழி ஆய்வுக்கு பயன்படுத்த முடியாது. குறிப்பாக இயந்திர மொழிப்பெயர்ப்புக்குச் சொல்வகைப்பாடு அடையாளப்படுத்துதல் மிக முக்கியமானதாகும்.

தமிழ்த் தரவுத் தொகுதிகளை சொல்வகைப்பட்டிறிகு அடையாளப் படுத்தல் பற்றி குறிப்பாக இந்திய மொழியிலிருந்து இந்திய மொழிகளின் மொழிபெயர்ப்பு ஒழுங்குமுறையில் தமிழ்த் தரவுத் தொகுதி எவ்வாறு அடையாளப்படுத்தப்படுகின்றன எச்சொல்வகை அடையாளங்கள் பயன் படுத்தப்படுகின்றன என்பன குறித்து இங்கு விரிவாக விளக்கப்படும்.

4.2 தரவுத்தொகுதியை அடையாளப்படுத்தல்

தரவுத்தொகுதி பல மொழியியல் தகவல்களுக்காக அடையாளப் படுத்தப்படுகின்றன. சொல்வகைப்பாடு அடையாளப்படுத்தல் (Parts of Speech Annotation), மீக்கூறு அடையாளப்படுத்தல் (Prosodic Annotation), முன்வரு கிளவி அடையாளப்படுத்தல் (Anaphoric Annotation), பொருண்மை அடையாளப்படுத்தல் (Semantic Annotation), கருத்தாடல் அடையாளப்படுத்தல் (Discourse Annotation)

என்பன முக்கியமான தரவுத்தொகுதி அடையாளப்படுத்தல்களாகும் (Annotation of corpus).

அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆய்வுக்கு மிகப் பயனுள்ளதாக அமையும். இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி சொற்கள் சொல்வகுப்புகளாக அடையாளப்படுத்தப்பட்டிருக்கும் தரவின் மிகப் பொதுவான வடிவாகும். Brown Corpus, Lorcorpus, BNC என்பவை இலக்கணத்திற்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதிகள் ஆகும். LLC (Lond Lund Corpus of Spoken English) மீக்கூறுகளுக்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதியாகும். Suffanne corpus தொடரியலுக்காக அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி ஆகும்.

4.2.1 சொல்வகைப்பாடு அடையாளப்படுத்தல்

சொல்வகைப்பாடு அடையாளப்படுத்தலின் (Parts of Speech Annotation) நோக்கம் ஒரு பனுவலில் உள்ள ஒவ்வொரு சொல் அலகுக்கும் அதன் வகைப்பாட்டைக் காட்டும் ஒரு கூறிலே தருவதாகும். இது விரிதரவிலிருந்து தரவை மீட்டெடுப்பதில் தனித் தன்மையைக் கூட்டுகிறது. இது தொடரியல் பகுத்து அடையாளப்படுத்தலுக்கும், பொருண்மைக் கள அடையாளப்படுத்தலுக்கும் உதவுகிறது. இது ஒற்றுமையான எழுத்துக்களை வேறுபடுத்துவதற்கு நம்மை அனுமதிக்கின்றது. சொல்வகை அடையாளப்படுத்தல் பின்வரும் தன்மைகள் கொண்டதாகும்:

- ஒவ்வொரு சொல்லுக்கும் ஒரு சொல்வகைப்பாட்டை அல்லது பிற சொல்வகை அடையாளத்தைத் தருவதாகும்.
- அடையாளப் படுத்துகைப் பொதுவாக நிறுத்தற்குறிகளுக்கும் பயன்படுத்தப்படும்.

- கணினிமொழியில் அடையாளப் படுத்துதல் [tokenization] போல் தான் இயற்கை மொழியிலும் சொல்லை அடையாளப்படுத்தும் செயல்பாடு அமையும்.
- சொல்வகை அடையாளப்படுத்தல் பேச்சுப் புரிந்து கொள்ளுதல் [Speech Recongition] இயற்கை மொழி பகுத்துக் குறித்தல் [Natural Language Passing]மற்றும் தகவல் மீட்டுப் பெறல் [Information retrival] என்பனவற்றில் முக்கியமான பங்கு வகிக்கின்றது.
- சொல்வகை அடையாளப்படுத்தலின் உள்ளீடு [Input] வழிமுறை வரைவு [algorithm] சொற்களின் கோர்வைகளும் குறிப்பிட்ட சொல்வகைப்பாட்டு அடையாளக் குழுமமும் [Part of Speech Tagsets] ஆகும்.
- விடுவரல் [Output] ஒவ்வொரு சொல்லுக்கும் ஒவ்வொரு நல்ல அடையாளத்தைத் தருவதுதான்.

  எ.கா:

  அவன் <மா.பெ> என் <வ.மா.பெ> சொல் <வி.பெ> படி <வி.பெ.பி.உ> கேட்டான்.

POS அடையாளப்படுத்தியின் முக்கியமான வேலை சூழலுக்குத் தக்கவாறு சரியான அடையாளத்தைத் தேர்ந்தெடுத்து மயக்கத்தைத் தீர்ப்பதுதான். பெரும்பாலான தமிழ்ச் சொற்கள் எல்லாம் மயக்கம் இல்லாதவை. இருப்பினும் பொதுவாகப் பயன்படுத்தப்படும் சொற்கள் மயக்கம் உள்ளவை.

4.2.2 முன்வருகிளவி அடையாளப்படுத்தல்

முன்வருகிளவி அடையாளப்படுத்தலில் (Annaphic Annotation) எல்லா மாற்றுப் பெயர்களும் பெயர்த்தொடர்களும் அகப்பற்றாக (Cohesion) என்பதன் விரிவான சட்டகத்திற்குள் ஒரே குறிபடு பொருளைக் குறிப்பதற்காக

அடையாளப்படுத்தப்படுகிறது. இங்கு முன்வருகிளவியும் வேறுபட்ட வகைகள் வகைப்படுத்தப்பட்டு வரிசைப்படுத்தப்படுகின்றன. இந்த முன்வருகிளவி அடையாளப்படுத்தப்படுத்தல் திட்டம் மாற்றுப்பெயர் தீர்வு (Anaphora resolution) போன்ற ஆய்வு மற்றும் பரிசோதனைச் செய்யு[ம் இயக்க நெறிமுறைகளுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா.

ரவி தன் தங்கையைப் பார்க்கப்போகும் ராமுவைச் சந்தித்தான்

இது பனுவல் புரிந்து கொள்கைக்கும், இயந்திர மொழிபெயர்ப்புக்கும் முக்கியமானதாகும்.

4.2.3 மீக்கூறு அடையாளப்படுத்தல்

மீக்கூறு அடையாளப்படுத்தலின் (Prosodic Annotation) இசையோட்ட ஒழுங்மைப்புகள், அழுத்தம் மற்றும் விட்டிசை என்பன பேச்சில் அடையாளப் படுத்தப்படும் இது மிகச் சிரமமான அடையாளப்படுத்தலே ஆகும். ஏனெனில் மீக்கூறு பிறமொழி நிலைகளை விட இயல்பில் தன் உணர்ச்சி அடிப்படையில் அமையும். இது பயிற்சி பெற்ற காதால் கவனமாகக் கேட்கப்படுத்தலைக் கேட்கும் Lancaster/IBM Spoken English Corpus மீக்கூறுக்காக அடையாளப்படுத்தப்பட்ட விரிதரவாகும். இதில் அழுத்தப்பட்ட அசைகள் சுதந்திரமான இசைமை இயக்ககத்துடன் மற்றும் இயக்கம் இல்லாமல் வேறுபட்ட குறியீடுகளால் அடையாளப்படுத்தப் பட்டுள்ளன. இசைமையைச் சூழலிலிருந்து ஊகிக்க இயலும். அழுத்தம் இல்லாத அசைகள் அடையாளப்படுத்தப்படாமல் விடப்பட்டுள்ளன.

4.2.4 பொருண்மை அடையாளப்படுத்தல்

பொருண்மை அடையாளப்படுத்தலில் (Semantic Annotation) பனுவலில் உள்ள சொற்களின் பொருண்மைப் பண்புக் கூறுகளோ (குறிப்பாக சொல் அர்த்தத்தின் அடையாளப்படுத்துகை) அல்லது சொற்களுக்கு இடையேயுள்ள பொருண்மை உறவுகளோ (குறிப்பிட்ட செயலின் செயல்கள், பாதிப்பவர்கள் போன்றவை) அடையாளப்படுத்தப்படுகின்றது.

எ.கா.

அவள் பழம் சாப்பிட்டாள்

அவள் - (படர்க்கை பெண்பால், ஒருமை, மாற்றுப்பெயர்,பதிலீடுபெயர்

பழம் - (தாவரத்தின் பூவிலிருந்து விளையும் விதையுள்ள இனிப்பான பொருள்

சாப்பிட்டாள் - (பழத்தை உட்கொள்ளுதல்)

எந்தப் பொருண்மைப் பண்புக் கூறுகள் அடையாளப்படுத்தப்பட வேண்டும் என்பதில் ஒருங்கிணைந்த கருத்து இல்லை. சிலர் Roget's Thesaurus –இல் உள்ள பொதுமையான பொருண்மை வகைப்பாடுகளை முன் வைத்தனர். இம்மாதிரியான அடையாளப்படுத்தும் திட்டம் சொற்களின் மூடப்பட்ட வகுப்புகளையும் திறந்த வகுப்புகளையும் நடைமுறைப்படுத்த அடையாளப்படுத்தப் படுகின்றது.

4.2.5 கருத்தாடல் அடையாளப்படுத்தல்

கருத்தாடல் அடையாளப்படுத்தலில் (Discourse Annotation) ஒரு தரவுத்தொகுதி பனுவல் மற்றும் கருத்தாடல் நிலைகளில் அடையாளப்படுத்தப்படுகின்றது. அடையாளப்படுத்துகின்ற வகை கருத்தாடல் ஆய்வில் முக்கியப் பங்களிப்பைச் செய்தாலும் பரவலாகப் பயன்படுத்தப்படுவதில்லை. ஏனெனில் மொழி வகைப்பாடுகள் சூழல் சார்ந்தவை மற்றும் அவற்றைப் பனுவல்கிளல் அடையாளம் காணுதல் பிற மொழியியல் நடப்புகளைக் காட்டிலும் விவாதத்திற்கு கூடுதல் மூல காரணமாக அமையும்.

Londond – Lund – Spoken Corpus–இல் வயதுக்கு வந்தவர்களின் பேச்சு போக்கைக் கவனிக்க 16 கருத்தாடல்கள் பயன்படுகின்றது.

4.2.6 பகுத்துக் குறித்தல்

பகுத்துக்குறித்தல் என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்க ஆய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும். வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுத்துக் குறித்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளில் கொண்டு வருகிறது. வாக்கிய நிலைப் பகுத்துக் குறித்தல் சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்துக் குறித்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree bank) என்றழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா:

'Pearl sat on a chair' என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – sat VVI] [PP on II]
[NP a – ATI] [chair – NNI]
[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்துக் குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.
2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பதுத்துக் குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்துக் குறித்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்துக் குறித்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகின்றது. பகுத்துக் குறித்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்துக் குறித்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்துக் குறித்தலின் குறைப்பாடு பகுத்துக் குறித்தல், தரவுத்தொகுதி திரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள்

சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

4.2.7 தலைச்சொல்லாக்கம்

தலைச்சொல்லாக்கம் (Lemmatization) செயல்பாடு விரிதரவில் பயன்படுத்தப்பட்டுள்ள சொற்களின் மூல வகைப்பாட்டைக் கண்டுபிடிப்பதும் சொற்களை அவற்றுடன் தொடர்புடைய சொல்லன்களாகக் (Lexeme) குறைப்பதும் ஆகும். எடுத்துக்காட்டாக *படித்தான்*, *படிக்கிறான்*, *படிப்பான்*, *படிப்பாள்* என்ற சொல் வடிவுகளிலிருந்து *படி* எனக் குறைத்தல். இது சொல்லின் எல்லா மாற்று வடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச் சொல்லின் மாற்றுவடிவங்களையும் உள்ளீடு செய்யாமல் ஒரு தலைச்சொல்லின் மாற்று வடிவங்களைப் பிரித்தெடுக்கவும், பரிசோதிக்கவும் ஆய்வாளர்களை அனுமதிக்கின்றது. இது மொழி கற்றலுக்குப் பயன் உள்ளதாக அமைகின்றது. இங்கு கற்பவர்கள் ஒரு தலைப்புச் சொல்லின் மொத்த சாத்தியமான எண்ணிக்கையில் பயிற்சி பெறுகின்றார்கள். இது எந்தச் சொற்கள் திரிபுறுகின்றன என்பதை அறிவதற்குப் பயன் உள்ளதாக இருக்கின்றது. எடுத்துக்காட்டாக Brown Corpus–இன் ஒரு பகுதி சொல் மற்றும் இலக்கணத் தகவலுடன் கூடிய சொற்களின் தலைப்பாக்கம் செய்யப்பட்ட வடிவுகளைக் கொண்டிருக்கின்றது.

4.3 சொல்வகை அடையாளப்படுத்தலின் முக்கியத்துவம்

மொழி ஆய்வில் சொல் வகைப் பாட்டின் முக்கியத்துவம் என்னவென்றால் இது சொல்லைப் பற்றியும் அவற்றை அடுத்து வரும் சொற்கள் பற்றியும் செய்திகளை ஏராளமாகத் தருகின்றது. இவை முக்கிய வகைப்பாடுகளுக்கு உண்மையாகும். வினை-பெயர் மட்டுமின்றி பல நுட்பமான வேறுபாடுகளுக்கும் உண்மையாகும். எடுத்துக்காட்டாக இந்த அடையாளக் குழுமங்கள் உடமை மற்றும் பெயர்களுக்கும் (*என்னுடைய, உன்னுடைய, அவனுடைய, அவளுடைய, அவருடைய, அதனுடைய*) மூவிடப் பெயர்களுக்கும் (*நான், நீ, அவன், அவள், அது*) வேறுபாட்டைக் காட்டுகின்றன.

ஒரு சொல் உடமை மாற்றுப்பெயரா? அல்லது மூவிடப்பெயரா? என்பதை அறிந்துகொண்டு எந்தச் சொற்கள் அதன் பக்கத்தில் வர இயலும் என்பதைக் கூற இயலும். இதைப் பேச்சுத் தெரிந்துகொள்ளுதல் (Speech Recognition) என்பதற்கான மொழி மாதிரியில் (Language Model) பயன்படுத்தலாம். சொல்வகைப்பாடு தகவல் மீட்டலுக்கான (Information retrieval) வேர்ச்சொல்லைக் கண்டுபிடித்தலுக்குப் பயன்படுகின்றது.

ஏனென்றால் சொல்வகைப்பாட்டை (Parts of Speech) அறிவது அவை எந்த உருபனியல் ஒட்டுகளையும் ஏற்கும் என்பதை அறியலாம். அவை பனுவலிலிருந்து பெயர்களையோ முக்கிய சொற்களையோ அறிய உதவுவதால் தகவல் மீட்டல் பயன்பாட்டிலும் பயன்படுத்தலாம். தானியங்கு சொல்வகைப்பாட்டு அடையாளப் படுத்திகள் சொல்பொருள் மயக்கம் நீக்கி வழியமைப்பு முறை (word sense disambiguation algorithms) சொல் வகைப்பாட்டு அடையாளப்படுத்திகள் (Part-of-speech-taggers=pos taggers), வகுப்பு அடையாளப்படுத்திகள் (class based N-grams) போன்ற முன்னேறிய ASR மொழி மாதிரிகளில் பயன்படுகின்றது.

ஆங்கிலத்தில் புகழ்பெற்ற POS tagger Brown Corpus-இல் பயன்படுத்தப்பட்ட 87 அடையாளக் குழுமங்களிலிருந்து (tagsets) உருவாக்கப்பட்டவை. மூன்று மிகப் பொதுவாகப் பயன்படுத்தப் படுகின்றன.

1. 45--Tag Penn Tree bank tagset.

2. The medium Size 61 tag 65 tagsets used by Lancaster VCREL, Projects CLAWS (Constituent Liklihood Audomatic Word-Tagging System) Tagger to tag British National Corpus (BNC)

3. The Longer 146-tag C7 tagset

Ex.

Penn Tree bank Version of Brown Corpus

ASCII Flat file

Tag represented after each word following a slash

The/DT grand /JJ/ Jury/NN commented/VBD on/IN a/DT number/NN of/IN others/JJ Topics/NNS%

DT – Deter miners

JJ – Adjective

NN – Noun,sing,or mass

NBD – Verb, Past tense

IN – Preposition

NNS – Noun plural

4.4 சொல் வகைப்பாட்டை அடையாளப்படுத்தலின் வகைகள்

மூல சொல்வகை அடையாளப்படுத்துவதின் வழிமுறை வரைவு (Algorithm) இரு வகுப்புகளில் அடங்கும். மூன்று அடையாளப்படுத்திகள் (taggers) பயன்பாட்டடில் உள்ளன. அவையாவன:

1. விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Rules based taggers)

2. புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochatic Taggers)

3. மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள் (Transformation based taggers)

4.4.1 விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள்

விதி அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் மயக்கம் தருகின்ற சொல் பெயரா?, வினையா? என்று குறிப்பிடுகையால் எழுதிய பெரிய தரவு மையங்களை [data base] உள்ளடக்கியது. ஆரம்பக் கட்டத்தில் தானாகவே சொல்வகைப்பாட்டை அடையாளப்படுத்துகின்றது. அவ்வாறு அடையாளப்படுத்த பயன்படுத்தப்பட்ட வழிமுறை வரைவுகள் (Algorithm) இரண்டு நிலை அமைப்பால் ஆனதாகும்.

1. அகராதி ஒவ்வொரு சொல்லுக்கும் சாத்தியமான சொல்வகைப்பாட்டைத் தரும்.

2. இரண்டாம் நிலை கையால் எழுதப்பட்ட மயக்கம் தரும் சொல்வகைப்பாட்டைத் தருமாறு பயன்படுத்தும்.

எடுத்துக்காட்டாக ENG Two Level Tagger இரண்டாம் நிலை அமைப்பால் ஆனதாகும். ENG TWO Level அகராதி இருநிலை உருபனியல் [Two level Morphology] அடிப்படையில் அமைந்தது.

4.4.2 புள்ளியியல் அடிப்படையிலான சொல்வகை அடையாளப்படுத்திகள் (Stochostic Part of Speech Tagging)

சாத்தியங்கள் அடையாளப்படுத்துவதில் பயன்படுத்துவது மிகப் பழமையானது. பல சாத்திய அடிப்படையிலான அடையாளப்படுத்திகள் வழக்கில் உள்ளன. எ.கா. Hidden Markove Model or HMM Tagger. இந்தப் புள்ளியியல் அடிப்படையிலான அடையாளப்படுத்தியில் பயன்படுத்தப்படும் உள்ளுணர்வு 'ஒரு சொல்லின் கூடுதலான சாத்தியமான அடையாளத்தை எடு' [Pick up the most likely tag for the Word] என்பதாகும். தரப்பட்ட வாக்கியத்திற்கு அல்லது சொல் கோர்வைக்கு HMM அடையாளப்படுத்தி பின்வரும் வாய்ப்பாட்டைப் (formula) பெரிதுபடுத்தும் அடையாளக் கோர்வையைத் தேர்ந்தெடுக்கின்றது.

P(wird/Tag) *P(Tag/Previous n tag)

HMM ஒரு சொல்லைக் காட்டிலும் ஒரு முழு வாக்கியத்திற்கு அடையாளக் கோர்வையைத் தருகின்றது. இவ்வகையான Bigram – HMM Tagger Wi என்ற சொல்லுக்கு ti என்ற அடையாளத்தைத் தேர்ந்தெடுக்கும். அது முந்தைய அடையாளம் ti–ஐ மற்றும் தற்போதைய சொல் wi என்று தரப்படுகையில் கூடுதலான சாத்தியமான அடையாளத்தைத் தேர்ந்தெடுக்கும்.

ti = arg max P(tj/ti-l,wi)

j

ஒரு அடையாளப் படுத்தலுக்கு அடிப்படையான HMM சமன்பாடு

ti argmax (tj,Pi-l) P(witj)

j

எ.கா:

(1) Secretariat,NNP is /VB2 expected/VBN to/To race VB Tomorrow/NN

(2) people/NNs continue NBP to/To inquire/VB The DT Reason/NN for/in The/CT race/NN For/in outer/JJ Space/NN.

முதல் எடுத்துக்காட்டில் race ஒரு வினை (VB) இரண்டாவது எடுத்துக்காட்டில் race ஒரு பெயர் (NN) Bigram HMM அடையாளப்படுத்தி அடையாளச் சிக்கலை அண்மையிலுள்ள சொற்களையும், அடையாளங்களையும் பார்த்துத் தீர்மானிக்கலாம் என்று ஊகம் செய்கின்றது.

To/To race/?????

Tge/DT race/????

எவ்வாறு race என்ற நமது எடுத்துக்காட்டு பயன்படுகின்றது என்று பார்ப்போம். நாம் NN -க்கும் VB -க்கும் இடையே ஒன்றைத் தேர்ந்தெடுக்க வேண்டும். நாம் எதற்குக் கூடுதல் சாத்தியம் இருக்கின்றதோ அதைத் தேர்ந்தெடுக்க வேண்டும்.

P(VB/To) P(race/VB)

P(NN/TO) P(race/NN)

பின்வருவது சாத்தியங்களாகும் (probabilities)

P(NN/TO) = :021

P(VB/To) = .34

4.4.3 மாற்றம் அடிப்படையிலான அடையாளப்படுத்திகள்

மாற்றம் அடிப்படையிலான அடையாளப்படுத்துதல் (Transformation Based Tagging) Bill Tagging என்று அழைக்கப் படுகின்றது. இது பொறி கற்றலுக்கான மாற்றம் அடிப்படையிலான கற்றல் (Transformation based Learning – TBL) அணுகுமுறையாகும். இது விதி அடிப்படையிலிருந்தும் புள்ளியியல் அடிப்படையிலிருந்தும் செய்திகளை எடுக்கின்றது. விதி அடிப்படையிலான அடையாளப்படுத்திகளைப் போல் TBL உம். எந்தச் சொற்களுக்கு எந்த

அடையாளத்தைத் தரவேண்டும் என்று குறிப்பிடுகின்றது. ஆனால் அடிப்படையிலான அடையாளப் படுத்திகளைப் போல் TBL ஒரு மொழி கற்றல் நுணுக்கமாகும் (Machine Learning Technique). இதில் தரவிலிருந்து விதிகள் தானாகவே ஊகிக்கப் படுகின்றது. சில HMM- Taggers -ஐ போல் TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் நுணுக்கமாகும் , (Supervised Learning Technique). இது ஏற்கெனவே அடையாளப் படுத்தப்பட்ட பயிற்சி பெற்ற பெருந்தரவை [ore-Tagged Training Corpus] எதிர்ப்பார்கின்றது,ஊகிக்கின்றது. TBL -உக்கு ஒரு குழுமம் அடையாளப்படுத்தும் விதிகள் இருக்கின்றது. முதலில் பெருந்தரவு பெரும்பாலும் பயன்படுத்தப்படும் விரிந்த விதிகளை வைத்து அடையாளப்படுத்தப்படும். பின்னர் சில முன்னதான [original] அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் தேர்ந்தெடுக்கப்படும். பின்னர் குறைவான எண்ணிக்கையிலான அடையாளங்களை மாற்றும். மேலும் திட்டவட்டமான விதிகள் பயன்படுத்தப்படும். எவ்வாறு TBL விதிகள் பயன்படுத்தப்படுகின்றது (How TBL Rules-are applied) என்பது கீழே விளக்கப்பட்டுள்ளது.

Brill's Tagger – ரை கொண்டு பயன்படுத்தப்படும் விதிகள் நடைமுறையைப் பயன்படுத்துவதற்கு முன் அடையாளப்படுத்தி ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும் இந்த கூடுதல் அடையாளங்களை நாம் பெருந்தரவில் பெறலாம்.

Brown Corpus –இல் 'race' என்பதற்குக் கூடுதல் உகந்தது பெயர் என்பதாகும்.

P(NN/race)=.98

P(VB/race)=.02

நாம் முன்னர் பார்த்த race என்பதன் இரு எடுத்துக்காட்டுகளும் NN என்று அடையாளப்படுத்தப்பட்டிருக்கும். முதல் எடுத்துக்காட்டில் NN என்பது தவறான அடையாளம்.

is/VB2 expected/VBN to/To race/NN tomorrow

இரண்டாவது எடுத்துக்காட்டில் race என்பது சரியாக NN என்று அடையாளப்படுத்தப்பட்டிருக்கும்.

The/DT race/NN for/IN outer/JJ space/NN

கூடுதல் உகந்த அடையாளத்தைத் தேர்ந்தெடுத்தப்பின் சேடைடள கூயபபநச அதன் மாற்ற விதிகளை நடைமுறைப் படுத்தும். இது நடக்கையில் race என்பதன் தவறான அடையாளப்படுத்தலில் நடைமுறைப்படுத்தப்படும் விதியைக் கற்றிருக்கும். முன்வரும் tag to என்றால் NN என்பதை VB என்றும் மாற்றுவதற்து முன் to/To என்று வந்தால் இந்த விதி race/NN என்பதை race/VB என்று மாற்றும்.

Expected/VBN to/To race/NN
Expected/VBN to/To race/VB

எவ்வாறு TBL விதிகள் கற்கப்படுகின்றன என்பது பின்வருமாறு விளக்கப்படும். Brill's TBL வழிமுறை வரைவுக்கு [algorithm] மூன்று முக்கிய நிலைகள் இருக்கின்றன. இது முதலில் ஒவ்வொரு சொல்லையும் கூடுதல் உகந்ததான அடையாளத்தால் அடையாளப்படுத்தும். பின்னர் இது ஒவ்வொரு சாத்தியமான மாற்றத்தையும் (Possible Transformation) பரிசோதிக்கும். இறுதியில் இருந்த விதிப்படி தரவை மீள் அடையாளப்படுத்தல் செய்யும். முன்னால் வரும் அடையாளப்படுத்தலை விட குறைந்த மேம்பாடுடைய அடையாளப்படுத்தலான

நிறுத்தல் அளவீடு வருவது வரை இந்த மூன்று நிலைகளும் திரும்பித்திரும்பி சொல்லின் சரியான அடையாளம் தெரியும் என்பதை வேண்டுகிறது. அதாவது TBL ஒரு கண்காணிக்கப்பட்ட கற்றல் வழிமுறை வரைவு (Supervised learning algorithm) ஆகும்.

TBL செயல்பாட்டின் வெளிப்யீடு முறைப்படுத்தப்பட்ட மாற்றங்களின் பட்டியலாகும். இவை புதிய பெருந்தரவுக்குப் பயன்படுத்தத்தக்க,நடைமுறை படுத்தத்தக்க அடையாளப்படுத்தல் செயல் முறையை (tagging Procedure) உருவாக்கும். கொள்கை அடிப்படையில் சாத்தியமான மாற்றங்களின் குழுமம் எல்லையற்றதாகும். ஆனால் மாற்றங்கள் ஒவ்வொன்றும் வழிமுறை வரைவைக் கடக்குமாறு மிகப் பொருத்தமான ஒன்றைத் தேர்ந்தெடுக்க TBL ஒவ்வொரு மாற்றத்தையும் எடுத்துக்கொள்ள வேண்டும். இவ்வாறு மாற்றங்களின் குழுமத்தை எல்லைப்படுத்த வழிமுறை வரைவுக்கு ஒரு வழி தேவை. இது சிறிய templates -களின் குழுமத்தை (உருமான மாற்றங்களை) உருவாக்கிச் செய்யப்படுகின்றது. பின்வருவன Brill's Templetes -களின் குழுமத்தின் பட்டியலாகும். முன்னர் வருகிற (தொடர்நது வருகிற) சொல் Z-ஆக அடையாளப்படுத்தப்படும். இரண்டு சொற்களுக்கு முன்னால் (பின்னர்) வருகிற சொல் Z -ஆக அடையாளப்படுத்தப்படும்.

முன்னர் வருகிற இரு சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற மூன்று சொற்களில் ஒன்று Z -ஆக அடையாளப் படுத்தப்படும். முன்னர் வருகிற சொல் Z -ஆகவும் இரு சொற்களுக்கு முன்னால் (பின்னால்) வரும் சொல் W -ஆகவும் அடையாளப்படுத்தப்படும்.

4.5 சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம்

4.5.1 சொல்வகை அடையாளக் குழுமங்கள் பற்றிய முந்தைய முயற்சிகள்

4.5.1.1 தொடக்ககாலச் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் உருவாக்கம் [1980-க்கு முன்]

1960-களிலும் 1970-களிலும் ஆங்கில மொழிக்கு வேண்டி சொல்வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கும் செயல்பாடு ஐக்கிய அமெரிக்காவில் தொடங்கப்பட்டு விட்டது. ஆரம்பக் காலத்தின் முக்கியமான சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் Klein – Simmons(1963) மற்றும் Gree – Fubin (1971) என்பவர்களால் உருவாக்கப்பட்டது. மற்றொரு முயற்சி பென்சிலியா பல்கலைக்கழகத்தில் மேற்கொள்ளப்பட்டது.

4.5.1.2 ஆங்கில சொல் வகைப்பாட்டுக் குழுமங்கள் (1980-க்குப் பிறகு)

1980-1990-களில் லங்காஸ்டர் பல்கலைக்கழகத்தில் ஆங்கிலத்திற்கு வேண்டி ஒரு தொடர்ச்சியான சொல்வகைப்பாட்டு குழுமங்கள் CLAWS என்ற சொல்வகைப்பாட்டை அடையாளப்படுத்தும் மென்பொருளை பயன்படுத்தி உருவாக்கப்பட்டது. CLAWS என்ற சொல்வகைப்பாட்டு அடையாளக் குழுமம் LOB தரவுத்தொகுதிக்குப் பயன்படுத்தப்பட்டது. பின்னர் CLAWS -ன் C5 மற்றும் C7 சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் BNC விரிதரவை அடையாளப் படுத்தப் பயன்படுத்தப்பட்டது. இதைத் தொடர்ந்து பல ஆங்கில சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் பயன்படுத்தப்பட்டன. அவையாவன:

1.The TOSCA Scheme

2.The ICE Tagset

3.The PENN Tagset

4.The LUND Tagset

5.The Eng CG Tagset

4.5.2 அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பு

ஆங்கிலம் தவிர பிற மொழிகளுக்குச் சொல் வகைப்பாடு அடையாளம் செய்யும் போது சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்தும் தேவை வெளிப்பட்டது. சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் மிக சமீபயத்திய தரமான ஒன்று EAGLES—வழிகாட்டல்களாகும் (The ENGLES-Guidelines) [Leeach and Wilson] இதனுடைய நோக்கம் பலவிதமான திட்டங்களுக்கும் பல மொழிகளுக்கும் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களைத் தரப்படுத்துவதாகும்.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உள்ளடக்கும் அல்லது உள்ளடக்க வேண்டிய ஒரு குழும பண்புக்கூறுகளைத் தருகிறது. இதனுடன் இந்த எல்லாப் பண்புக்கூறுகளையும் குறியாக்கம் செய்து இடைநிலைப்பட்ட குறியாக்கத்தின் தேவை என்னவென்றால் இரண்டு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கு இடையில் மொழிப்பபெயர்ப்பை அனுமதிப்பதாகும். இதனால் அவற்றிற்கிடையில் பொருத்தத்தை உறுதி செய்வதாகும். உருப-தொடரியல் அடையாளப்படுத்தலுக்கு EAGLES வழிகாட்டல்களில் மூன்று காட்டல்களில் அமைப்பாக்கம் செய்யப்பட்டுள்ளது.

1. எவை எவற்றைக் கட்டாயமாகக் கருதப்படுகிறது.
2. எவை பரிந்துரைக்கப்படுகின்றது.
3. எவை உருப-தொடரியல் அடையாளப் படுத்தலுக்கு சின்னம் - மதிப்பிணைகளாக வரையறை விளக்கம் செய்யப்படுகிறது. எடுத்துக்காட்டாக 'பால்' என்ற சின்னத்திற்கு EAGLES -இன் பரிந்துரைகளில் ஆண்,பெண் அல்லது பால் மதிப்பிணைகள் படிநிலையாக அமைக்கப்படவோ அமைக்கப்படாமலோ இருக்கலாம்.

EAGLES வழிகாட்டல்களால் எந்தச் சொல் வகைப்பாட்டு அடையாளக் குழுமங்களுக்கும் கட்டாயமாக பதிமூன்று பரிந்துரைக்கப்படுகின்றது.

1. Noun (பெயர்)
2. Verb (வினை)
3. Adjective (பெயரடை)
4. Pronoun (பதிலீடு பெயர்)
5. Article (சுட்டிடைச்சொல்)
6. Adposition (உருபு)
7. Adverb (வினையடை)
8. Conjunction (இணைப்பான்கள்)
9. Numeral (எண்:
10. Interjection (வியப்பிடைச்சொல்)
11. Unassigned/unique (இணையற்ற,தனிப்பட்ட)
12. Residual (மீதியான)
13. Punctutation (நிறுத்தற்குறிகள்)

பரிந்துரைக்கப்பட்ட மற்றும் இருப்புச் சின்னங்கள் இந்த முக்கியமான சொல் வகைப்பாடுகளால் ஒழுங்குப்படுத்தப் பட்டுள்ளது. சொல் வகைகளை கடந்து பொருத்தமுறத் தேவையில்லை.

EAGLES வழிகாட்டல்கள் சொல் வகைப்பாட்டு அடையாளக் குடும்பத்தைத் திட்டமிடுபவரின் சுதந்திரத்தைக் கட்டுப்படுத்தாமல் அவர் குறிப்பிட்ட எண்ணும் எல்லா அடிப்படையான செய்திகளையும் உள்ளடக்கி நிகழ இயலும் சட்டகத்தைத் தருகின்றது. முன்னர் விளக்கப்பட்ட பல அடையாளக் குழுமங்கள் EAGLES வழிகாட்டல்களில் எளிதில் பொருந்தாத பண்புக் கூறுகளை உள்ளடக்கலாம். இருப்பினும் இவைகள் மற்றொரு அடையாளக் குழுமத்திற்கு மொழிப்பெயர்க்க இயலாத குழுமமும் அவற்றை உள்ளடக்கவில்லை. எனவே அவற்றை EAGLES

இடைப்பட்ட அடையாளக் குழுமத்தில் அடக்குவது தேவையில்லாததாகும். இந்தச் சட்டகத்தின் மதிப்பு இது வேறுபட்ட மொழிகளின் மொழியியல் வளங்களின் திண்மையையும் திரும்ப பயன்படுத்தலையும் ஊக்குவிக்கின்றது. மற்றும் சுழலான திரும்பக் கண்டுப்பிடித்தலை ஊக்கமிழக்கச் செய்கிறது.

EAGLES வழிகாட்டலின் முக்கியமான குறை இது. உலக மொழிகளின் மிகக் குறுகிய பகுதியைதான் அடக்குகிறது. ஐரோப்பிய ஒருங்கிணைப்பின் திட்டமாக இது ஆங்கிலம், டச்சு, ஜெர்மனி, டேனிக்ஷ், பிரெஞ்சு, ஸ்பேனிக்ஷ், போர்த்துகீசு, கிரீக், இட்டாலின் எனற வகைப்பாட்டியல் அடிப்படையில் ஒற்றுமையுள்ள ஒன்பது மொழிகளைத் தான் உள்ளடக்குகிறது. எனவே இவை அல்லாத பிறமொழிகளுக்கு நுஆழுடுநுளு வழிகாட்டல்களின் பயன்பாடு பொருத்தமாகவோ, பொருத்தமற்றதாகவோ இருக்கலாம்.

4.5.3 EAGLES வழிகாட்டல்களின் அடிப்படையில் உருவாக்கப்பட்ட சில சமீபக் காலத்திய சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

EAGLES வழிகாட்டலின் வெளிப்யீட்டிற்குப் பிறகு அவற்றைப் பயன்படுத்தி பலவிதமான மொழிகளுக்குச சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள் உருவாக்கப்பட்டன. அவை பின் வருமாறு:-

1.The MULTEXT 20 Project

2.The GRACE Project

3.The CRATER Project

EAGLES வழிகாட்டல்களால் அடக்கப்படாத மொழிகளின் சமீப காலத்திய சொல்வகைப்பாட்டு அடையாளக் குழுமங்களில் பின்வருவன அடங்கும்.

1.Tagset Design for Arabic.

2.Tagset Design for Chinese.

3.Tagset Design for Korean.

4.Tagset Design in the Paninijan Tradition.

4.6. தமிழ்த் தரவுத்தொகுதிகளை சொல்வகைப்பாட்டிற்கு அடையாளப்படுத்தல்

4.6.1. தமிழ்ச் சொல்வகைப்பாடுகள்

ஒரு மொழியின் சொற்கள் சொல்லியல் அடிப்படையிலும், தொடரியல் அடிப்படையிலும், வகைப்பாடு செய்யப்படலாம். சொற்கள் எந்த இலக்கணக் கூறுக்காகத் திரிபுறும் வாக்கியத்தில் எங்கு, எவ்வாறு இடம் பெறுகின்றன. அவற்றின் செயற்பாடுகள் யாவை என்பதன் அடிப்படையில் அவை ஒரு சொல் வகுப்பைச் சார்ந்தன என நிர்ணயிக்கலாம். ஆங்கிலத்தில் “Part of speech tag” என்றும்,இதன் படி சொற்களை எட்டு வகுப்புகளாகப் பிரிப்பர்.

1. பெயர் - Noun
2. பெயரடை - Adjective
3. மாற்றுப்பெயர் - Pronoun
4. வினை - Verb
5. வினையடை - Adverb
6. முன்னுருபு - Proposition
7. இணைப்புக்கிளவி - Conjunction
8. வியப்புச்சொல் - Interjection

தற்காலத் தமிழ் இலக்கணத்தை ஆங்கில இலக்கண கண்ணோட்டத்தோடு நோக்கும் இலக்கண நூல்களில் ஆங்கிலச் சொல்வகைப்பாட்டின் தாக்கத்தினைக் காணலாம். பழந்தமிழுக்காக எழுதப்பட்ட மரபிலக்கணங்களில் சொல் வகைகள்

தற்காலத் தமிழுக்காக மாற்றியமைக்கப்பட வேண்டும் என மொழியியல் அறிஞர்கள் கருதுகின்றனர். இருப்பினும் தற்காலத் தமிழுக்கு எத்தனை சொல் வகுப்புகள் நிர்ணயிக்கப்பட வேண்டும் என்பதில் அவர்களிடையே ஒருமித்த கருத்து இல்லை.

மரபிலக்கணங்கள் தமிழ்ச் சொற்களை பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், உரிச்சொல் என நான்கு வகுப்புகளாக கொண்டு, பெயரும்,வினையும் முக்கியமானதாகவும் இடையும்,உரியும் துணை வகுப்புகளாகவும் கொள்ளப்படுகின்றது. காலம் காட்டாமல் வேற்றுமை உருபுகளை ஏற்கும் சொற்கள் பெயர்ச் சொற்களாகவும், வேற்றுமை உருபுகளை ஏற்காது காலம் காட்டும் சொற்களை வினைச்சொற்களாகவும் கொண்டுள்ளனர். தனித்து இயங்கும் ஆற்றலின்றிப் பெயர்களையும்,வினைகளையும் இடமாகக் கொண்டு அவற்றின் முன்னும், பின்னும் வரும் வேற்றுமை உருபுகள், விகுதிகள், சுட்டெழுத்துகள், வினாவெழுத்துக்கள் ஆகியனவும் இடைச்சொல் எனப்பட்டன. பெயர்ச்சொல், வினைச்சொல் போல் தனித்து நின்று பொருளை உணர்த்தாமல் பெயர்ச்சொல்லையோ, வினைச்சொல்லையோ சார்ந்து நின்று அவற்றின் குணத்தை “மிகுதி” என உணர்த்த வல்லன உரிச்சொல்.

மரபிலக்கணத்தார் சொற்பகுப்பை ஆராய்ந்தால் சொற்களைப் பெயர்ச்சொல், வினைச்சொல் என்ற இரு பெரும் பிரிவுகளாகவும், அவற்றைச் சார்ந்து வரும் மூன்றாவது பிரிவுச்சொற்களை இடைச்சொல்,உரிச்சொல் என்று பிரித்தனர். மரபிலக்கணங்கள் உரிச்சொற்களைப் பெயர், வினை இவற்றிற்கு அடையாகக் கொள்கின்றனர். இவ்வாறு மரபிலக்கணங்கள் பெயர்ச்சொல், வினைச்சொல், இடைச்சொல், அடைச்சொல் என நான்கு வகைச் சொல் வகைகளை அடையாளம் காட்டுகின்றன. முற்காலத்தமிழை மரபிலக்கண வழி ஆய்கையில் இந்நான்கு வகை சொல் பகுப்புகள் போதுமானதாக இருந்தது.

தற்காலத் தமிழுக்கான சொல்வகுப்புகள் பற்றி இராசேந்திரன் அவர்கள் தமது கட்டுரையில் விரிவாகப் பேசுகிறார். அவர்கருத்துகள் இங்குத் தொகுக்கப்பட்டுத் தரப்பப்டுகின்றன். தற்காலத் தமிழ் புதிய இலக்கணக் கூறுகளை தன்னகத்தே வளர்த்துக் கொண்டுள்ளமையால் சொற்களை புதிதாக வகைப்படுத்த மொழியியல் அறிஞர்கள் வாதிட்டும், வகைப்படுத்தியும் உள்ளனர். [Asher:1982, லேமன்:1989, Lehman:1989]

லேமன் சொல் வகைகளைப் பற்றி விரிவாக ஆராய்கிறார். [Lehman,1989, 911] தற்காலத் தமிழ் மொழியில் எல்லா அகராதி அல்லது வேர் உருபுகளை நான்கு வகைகளாகப் பிரித்து அவைகளை 1.பெயர் வேர்கள் 2. வினை வேர்கள் 3. வினையடை வேர்கள் 4. பெயரடை வேர்கள் எனக்கொள்கிறார். இவற்றில் பெரும்பாலான வேர்கள் பெயர், வினை வேர்களில் அடங்கும். பெயர்கள் பெயர் வேர்களையும், வினைகள் வினை வேர்களையும் கொண்டிருக்கும். ஆனால் பின்னுருபுகள் பல வினையடைகள், அளவையடைகள், இணைப்புச் சொற்கள் என்பனவற்றை வடிவ அடிப்படையில் பெயர் அல்லது வினை வேர்களின் திரிபுற்ற அல்லது திரிபுறா வடிவங்களாகக் கொள்ளலாம். பழந்தமிழில் சொல் வகுப்புகள் குறைந்த காரணத்தினால் பல திரிபுற்ற அல்லது திரிபுறாத பெயர், வினை மற்றும் பெயரடை வேர்கள் தற்காலத் தமிழில் பின்னுருபுகள், வினையடைகள், பெயரடைகள், அளவையடைகள் போன்ற பல சொல் வகுப்புகளாக மீள்ளாய்வு செய்யப்பெற்றன. நாம் அறிந்தபடி தற்காலத் தமிழில் இரண்டு சொல் வகுப்புகள் தான் திரிபுகளை ஏற்கவல்லது. அவை பெயரும் வினையும் ஆகும். பெயர்கள் வேற்றுமையுருபு, எண் இவற்றிறகாகத் திரிபுறும். வினைகள் காலம், எண், பால், இடம் மற்றும் சிலவற்றிற்காகவும் திரிபுறும். சொல் வகுப்புகளில் பெயர்கள் சொல்லாக்கத்தை காட்டுகின்றன. பெயர்களை வினைகளிலிருந்து பெறலாம்.

ஆஷர் [1982:101,102] ‘Operational definitions for word classes’ என்ற தலைப்பில் தமிழ்ச் சொற்களை ஆறு வகையாகப் பிரிக்கிறார்: 1. பெயர்ச்சொல்

(Noun), 2. மாற்றுப்பெயர் (Pronoun), 3. வினைச்சொல் (Verb), 4. பெயரடை (Adjective), 5. பின்னுருபு (Post position), எண், அளவையடை, 6. இடைச்சொற்கள் (Particles)

இதையே கோதண்டராமன் (1989) தனியே இயங்க வல்ல சொற்களைத தலைமை இலக்கணக் கூறுகளாகக் கொண்டு பத்து வகுப்புகளாகப் பிரிக்கிறார். அவையாவன: 1. பெயர், 2. வினை, 3. பெயரடை, 4. வினையடை, 5. வல்லடை, 6. இணைப்பான், 7. உணர்ச்சி வெடிப்புச்சொல், 8. உணர்வு குறிப்புச்சொல், 9. விளிப்புச்சொல், 10. விளி ஏற்புச்சொல்

மேலும் “விகுதி”, “ஒட்டு” முதலானச் சார்புநிலைக் கிளவிகளைத் துணைமை இலக்கணக் கூறுகளாகக் கொண்டு அவற்றின் “வடிவுக்கும் இயல்புக்கும் ஏற்ப” 1. விகுதி (Suffix), 2. சொல்லுருபு (postposition), 3. வினைசார் கிளவி (Verb Participle), 4. ஒட்டு (Clitic), 5. சாரியை அல்லது நிரப்பி (Fillers) என ஐந்தாக வகைப் படுத்துகிறார்.

லேமன்(1989:9) உருபனியல் மற்றும் தொடரியல் இலக்கண இயல்புகள் அடிப்படையில் தமிழ்ச்சொற்களை 1. பெயர்கள் (Nouns), 2. வினைகள் (Verbs), 3. பெயரடைகள் (Adjectives), 4. வினையடைகள் (Adverbs), 5. பின்னுருபுகள் (Post position), 6. அளவையடைகள் (qantifiers), 7. வரையறு அடைகள் (Determiners), இணைப்புக் கிளவிகள் (Conjunctions) என எட்டாக சொல் வகுப்புகளைப் பிரிக்கிறார்.

இவ்மூவறிஞர்களும் பெயர், வினை, பெயரடை என்ற சொல் வகுப்புகளைத் தேர்ந்தெடுப்பதில் உடன்படுகிறார்கள். கோதண்டராமனும், லேமனும் வினையடையைச் சொல்வகுப்பாகக் கொண்டு ஆக்ஷரிடமிருந்து வேறுபடுகின்றனர். இணைப்பான் என்று கோதண்டராமனும், இணைப்புக்கிளவி

என்று லேமனும் குறிப்பிடும் சொல்வகுப்பை ஆஷர் சொல்வகுப்பாகக் கொள்ளவில்லை. ஆஷரும், லேமனும் பின்னுருபுகளைச் சொல்வகுப்பாகக் கொள்ள கோதண்டராமன் பின்னுருபுகளைச் சொல்லுருபுகள் எனக்குறிப்பிட்டு சார்நிலைக்கிளவியாக வகைப்படுத்துகிறார்.

பெயர்ச்சொல்

ஆஷர் பெயர்ச்சொல் பிரிவை உருபனியல் மற்றும் தொடரியல் அடிப்படையில் சொல் வகுப்பாக நிறுவுகிறார். பெயர்ச்சொற்கள் உருபனியல் அடிப்படையில் வேற்றுமை ஒட்டுகளையும், பன்மை ஒட்டு "கள்" என்பதையும் ஏற்கும். தொடரியல் அடிப்படையில் பெயர்கள் பின்னுருபுத்தொடரின் தலைச் சொல்லாகவும், ஒரு வாக்கியத்தின் எழுவாயாகவோ, செயப்படு பொருளாகவோ செயல்படும். எழுவாயாக வரும் பெயர்ச்சொல்லின் இடம், எண், பாலுக்கு தகுந்தாற்போல் வினையின் இடம், எண், பால், ஒட்டின் தேர்வும் அமையும். வேற்றுமை உருபுகளையும் ஆக,ஆன கிளவிகளை ஏற்பனவும், எழுவாயாகவும், பயனிலையாகவும் செயல்பட வல்லன பெயர்கள் என்று கோதண்டராமன் குறிப்பிடுகிறார். லேமன் வேற்றுமை பின்னொட்டுகளையும் ஆக,ஆய் பின்னொட்டுகளையும் ஏற்பன பெயர்கள் என்று விளக்குகிறார். மேலும் குறை உருபனில் உள்ள பல பெயர்கள் எல்லா வேற்றுமை பின்னொட்டுக்களையும் ஏற்காது என்பதையும் சுட்டுகிறார். பெயர்ச்சொல் வகுப்பில் மூவருக்கும் அடிப்படை பேதம் இருப்பதாகத் தெரியவில்லை. லேமன் மாற்றுப்பெயர், அளவுப்பெயர், எண்ணுப்பெயர் இவற்றை பெயர்ச்சொல்லின் துணைப்பிரிவாகக் கொள்கிறார்.

மாற்றுப்பெயர்

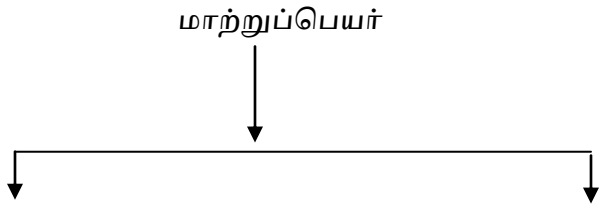
பெயர்ச்சொல்லுக்குப் பதிலாக அல்லது மாறாக வரும் சொல் மாற்றுப்பெயர் எனப்படும். இப்பெயர் முக்கியமாகப் பொருட்குறிப்பு (Signification)

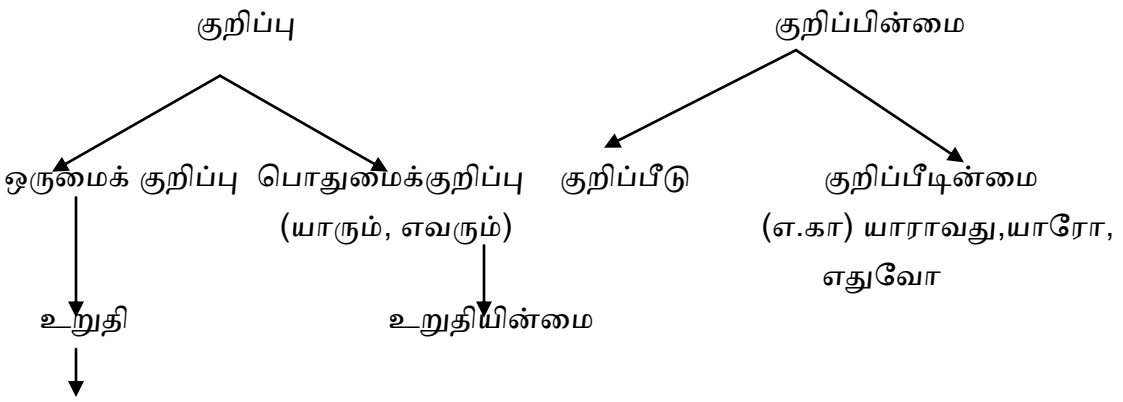
அடிப்படையில் பெயரிலிருந்து வேறுப்படும். ஒன்றையோ, பெயர்ச்சொல்லையோ குறிக்கும். ஆனால் மாற்றுப்பெயருக்குத் தக்கவாறு வெவ்வேறு ஆட்களையோ, பொருளையோ குறிக்கும். இதன் காரணமாய் ஒரு மொழியின் பெயர்ச்சொல்லின் எண்ணிக்கை ஏராளமாயிருந்தாலும், மாற்றுப்பெயர்கள் குறைவாகவே உள்ளது.

ஆஷர் மாற்றுப்பெயர்களைத் தனிச்சொல் வகுப்பாகக் கையாள்கிறார். பெயர் சொல்லின் வரையறைகளில் பல இயல்புகளைக் கொண்ட ஒரு நெருங்கிய குழுமத்தைச் சார்ந்திருப்பவை மாற்றுப்பெயர்கள். பெயர்ச்சொற்கள் ஏற்கும் அதே வேற்றுமைப் பின்னொட்டுகளை அவற்றுடன் சேர்க்கலாம். ஆவை பின்னுருபுத் தொடரின் தலைச் சொல்லாக அல்லது எழுவாயகவும், செயப்படுப்பொருளாகவும் செயற்படலாம். ஒரு வாக்கிய அமைப்பில் வினையின் இடம், எண், பால் ஒட்டைத் தேர்வு செய்யும்.

லேமன் பெயர்த் தொடரில் அடை எடுக்காது தனித்து வரும் பெயர்களை மாற்றுப் பெயர்களாகக் கொள்கிறார். மாற்றுப்பெயர்களைத் தனிநிலை மாற்றுப்பெயர்கள், [எ.கா.நான், அவன்] , ஆக்க மாற்றுப்பெயர்கள் [எ.கா.யாரோ, யாரும்] என்றும் வகைச் செய்கிறார். மூவிடப்பெயர்கள் [Personal Pronoun], சுட்டுப்பெயர்கள் [Demonstrative Pronouns], வினாப்பெயர்கள் [interrogative Pronouns], தன்வயப்பெயர்கள் [Reflexive Pronoun] என பலவகையாகப் பிரிக்கப்படுகிறது. மூலவகையான மாற்றுப்பெயர்களைக் குறிப்பு (Referential), உறுதி (Definiteness), குறிப்பீடு (Specificity) என்ற பொருண்மைக் கருத்துகள் அடிப்படையில் வனைப்பாடு செய்யலாம்.

லேமன் ,1989:93





மூவிடப்பெயர்கள் காலத்தோடும், இடத்தோடும் தொடர்பு கொண்டிருக்கும். பேசுபவசைத் தன்மை இடம் என்றும், கேட்பவரை முன்னிலை இடம் என்றும், பேசப்படுபவரைப் படர்க்கை இடம் என்றும் செய்யலாம். சுப்ல கூடுதலான பாகுபாடுகளைக் கொண்டு படர்க்கையிடம், தன்மை, முன்னிலை இடங்களிலிருந்து வேறுபடுகின்றது. தன்மை, முன்னிலை, படர்க்கை ஆகிய மூன்றும் அடிப்படையில் ஒருமை, பன்மை என வேறுபடுகின்றது. முன்னிலை இடங்களும், படர்க்கை இடங்களும் தகுதி அடிப்படையில் உயர்வு, உயர்வின்மை என வேறுபடும். படர்க்ளை இடப்பெயர்கள் மட்டுமே பால் அடிப்படையில் ஆண், பெண் என்றும், சுட்டு அடிப்படையில் அண்மை, சேய்மை என வேறுபடும்.

இதற்கான அமைப்பை கீழே அட்டவணை மூலம் காணலாம்

இடம்	ஒருமை	பன்மை
தன்மை	நான்	நாம்,நாங்கள்
முன்னிலை	நீ,நீர்	நீங்கள்
படர்க்கை	அவன்,இவன்,அவள்,இவள், அவர்,இவர்,அது,இது	அவர்கள்,இவர்கள் /அவை,அவைகள்

		இவை,இவைகள்

தான், நாம், தாங்கள் என்ற தன்வயப்பெயர்கள் நான்காவது இடத்தைச் சாரும். எழுவாயால் வரும் படர்க்கைச் சொல்லுக்கு மாறாய் அல்லது பதிலாய் தன்வயப்பெயர்கள் வரும்.

வினாப்பெயர்களுக்குத் தீர்மானிக்கப்பட்ட குறிப்பு இல்லை. வினாப்பெயர்களைக் குறிப்பீடு பெற்ற (Specified) வினாப்பெயர். குறிப்பீடு பெறாத (non specified) வினாப்பெயர் எனப்பிரிக்கலாம். அவை உயர்திணை, அஃறிணை வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. *யார்*, *என்ன*) குறிப்பீடுப்பெற்ற வினாப்பெயர்கள் படர்க்கையிடம் எண், பால் இவற்றின் வேறுபாட்டை வடிவத்தில் காட்டுகின்றன. (எ.கா. எவன், எவள், எவர், எது, எவை) உம் என்ற குறைச்சொல் சேர்ந்த வினாப்பெயர்களுக்கு எடுத்துக்காட்டுகளாகும். குறைச்சொல் ஒ-சேர்ந்த வினாப்பெயர்களான யாரோ, ஏதோ, என்னவோ என்பன குறிப்பீடு பெறாத உறுதியற்ற மாற்றுப்பெயர்களுக்கு எடுத்துக்காட்டுகளாய் அமையும்.

அளவுப்பெயர்

சில, பல, எல்லாம், எல்லோரும் என்ற பெயர்கள் அளவுப்பெயர்கள் பிரிவைச் சாரும். இவை பெயர் அடைகளாகச் செயல்பட வல்லன (எ.கா. சில மனிதர்கள்). பெயர்களுக்குப் பின்னால் வந்து வேற்றுமை ஒட்டுகள் எடுக்க வல்லன (எ.கா. பேனாக்கள் சிலவற்றை வாங்கினேன்). சில, பல என்பவைகளைப் பெயரடைகளாகக் கொண்டு முறையே சிலவை, பலவை என்பனவற்றை மாற்றுப் பெயரொட்டாக்கம் பெற்ற பெயர்களாக ஆயலாம் (எ.கா. சில+அவை=சிலவை)

எண்ணுப்பெயர்

எண்ணுப்பெயர்கள் ஆதார எண் (Cardinal Number), முறைமை எண் (Ordinal number) என இருவனைப்படும். முறைமை எண் ஆதார எண்ணிலிருந்து ஆவது அல்லது ஆம் என்ற குறைச்சொல்லின் (Clitics)சேர்க்கையால் ஆக்கப்படும். ஆதார எண்களிலிருந்து மாற்றுப் பெயரொட்டாக்கம் பெற்ற எண்ணுப்பெயர்களை (Pronominalzed Cardinal Number) பெறலாம். (எ.கா. ஒருவன், ஒருத்தி, ஒருவர்)

வினை

வினையைச் சொல்வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே கருத்து வேறுபாடு இல்லை. வினைகள் காலப்பின்னொட்டுகளையும், இடம் - எண் - பால் பின்னொட்டுகளையும் பிற பின்னொட்டுகளையும் எடுக்க வல்லன் சில வினைகள் குறை உருபனியலை உடையன. அதாவது எல்லா வினைப்பின்னொட்டுகளையும் அவை ஏற்பதில்லை. இணைப்பு வாக்கியங்கள் (Copular Sentences) (எ.கா. அவன் மாணவன்) தவிர்த்த பிற வாக்கியங்களில் கட்டாயமாக வரும் பகுதி வினையாகும். வினைகளை உருபனியல், பொருண்மையியல் இவற்றின் அடிப்படையில் வெவ்வேறு வகையாய் பாகுபாடு செய்யலாம். காலம் மற்றும் பிறவினைப் பின்னொட்டுகளுடன் சேரும்போது ஏற்படும் உருப ஒலியனியல் மாற்றங்கள் அடிப்படையில் வினைகளை மெல்வினை, இடைவினை, வல்வினை என்ற மூன்று வகைகளாகவும் அதற்கு மேற்பட்ட வனைகளாகவும் பிரிக்கலாம். வடிவ மற்றும் செயற்பாடு அடிப்படையில் வினையை முற்றுவினை (finite verb) (எ.கா. வந்தான்) மற்றும் எச்சவினை (non-finite berb) (எ.கா. வந்த, வந்து) என்று பிரிக்கலாம். எச்சவினை பெயரின் அடையாக வருமா அல்லது வினையின் அடையாக வருமா என்பதன் அடிப்படையில் பெயரெச்சம் (Adjectival or relative participle form) (எ.கா. வந்த பையன்) என்றும், வினையெச்சம் (adverbial or verbal participle form)

(எ.கா.வந்து போனான்) என்றும் வகைப்படுத்தலாம். தன்வினை (எ.கா.ஆடு) பிறவினை (எ.கா. ஆட்டு) என்ற பாகுபாடு பொருண்மையியல் அடிப்படையில் அமைந்தது. வினையினை தான் செய்யின் தன்வினை என்றும், பிறர் செய்யின் அது பிறவினை என்றும் விளக்கப்பட்ட பொருண்மை வேறுபாடு அவ்வளவு பொருத்தமானதல்ல. (பரமசிவம்,1983:2-3) அடிப்படையிலான வினைகளைத் தனவினைகள் என்றும், (எ.கா. ஓடு, காண்) அவற்றினின்றும் ஒட்டாக்கத்தால் ஆக்கப்பெற்ற வினைகளை (எ.கா. ஓட்டு, காட்டு) பிற வினைகள் என்றும் கொள்ளலாம். வாக்கியத்தில் பயன்படுத்துகையில் செயப்படுபொருளை எடுக்குமா? என்பதன் அடிப்படையில் வினைகள் செயப்படுபொருள் குன்றியவினை (எ.கா. போ, வா), செயப்படுபொருள் குன்றாவினை (எ.கா. படி, அடி) எனப்பாகுப்படுத்தலாம். பெயருக்கும் வினைக்கும் உள்ள காரக உறவுகளை (Case Relations)கொண்டும், இணைதிறன் (Valency))கொண்டும் வினைகளை வகைப்பாடு செய்யலாம்.

பின்னுருபுகள்

ஆக்ஷர் பின்னுருகைச் சொல்வகுப்புகளில் ஒன்றாகக் கூறுகின்றார். வேற்றுமைப் பின்னொட்டுகளால் திரிபுற்ற பெயர்களுக்குப் பின்வந்து பின்னுருபுத் தொடராக வினையுடன் ஒரு செயற்பாட்டு உறவில் நிற்கும் என்பது சொல் பின்னுருபு என்று தொடரியல் விளக்கம் தரலாம். பின்னுருபுகள் இந்நிலையில் தனித்து நிற்கும் என்பதிலிருந்து கட்டுப்பட்டே நிற்கும் என்பது பலதரப்பட்ட அங்கங்கள் சேர்ந்த சொல்வகுப்பாகும். கோதண்டசாமன்(1989) சாஹ்நிலைக் கிளவி வகை என்ற துணைமைச்சொல் வகையின் கீழ் பின்னுருபைச் சொல்லுருபு எனக் குறிப்பிட்டு வகைபாடு செய்கிறார். பெயர்ச்சொல்லும், வினைச்சொல்லும் காலப்போக்கில் வேற்றுமைப் பின்னொட்டுகள் போல் சொல்லுருபுகளாய் பயன்படுத்தப்படுகின்றன. லேமனும்(1989) பின்னுருபுகளைத் தனிச்சொல் வகுப்பாகக் கொள்கிறார். திரிபுற்ற மற்றும் திரிபுறாத பெயாவடிவுகளும் வினையெச்ச வடிவுகளும் தான் பின்னுருபுகளாகும். எடுத்துக்காட்டாக, தமிழ் பல

இடப்பொருள் செயற்பாடுகளைக் குறிப்பிட, இடத்தைக் குறிக்கும் பெயர்களைத் தான் பயன்படுத்துகின்றது. (எ.கா. இல், உள், முன், மேல், கீழ், அடியில்) ஒரு குறிப்பிட்ட சொல்லைப் பெயராகக் கருதுவதா, பின்னுருபாகக் கருதுவதா என்பதில் ஒருமித்த கருத்தில்லை. பின்னுருபுகளாக எடுத்தாளப்படும் பெயர்கள் எல்லாம் உருபனியல் அடிப்படையில் குறை உள்ளவை. அதாவது அவை எல்லா வேற்றுமை உருபுகளையும் எடுப்பதில்லை. பல பின்னுருபுகளாய் வரும் பெயர் வடிவுகள் தொடரியல் அடிப்படையில் குறைப்பாடுள்ளவை. அதாவது அவை வாக்கியத்தில் பெயர் வரும் எல்லா இடங்களிலும் வருவதில்லை. எல்லாப் பெயர்களைப் போலவும் அவற்றின் இலக்கணச் செயற்பாடுகள் வருவதில்லை. எடுத்துக்காட்டாக, நடு, இடை என்ற பெயர்கள் 'மையம்' என்ற பொருளில் இல் என்ற இடவேற்றுமை பின்னொட்டு எடுத்து நடுவில், இடையில் என பின்னுருபுகளாக மட்டும் வருகின்றன. ஆனால் அவை எழுவாயாகவோ, செயப்படுபொருளாகவோ, பயனிலையாகவோ வருவதில்லை. இப்பெயர் வடிவங்களில் மிகப் பெரும்பான்மையானவையாய் வரும் நெருங்கிய பெயர் மற்றும் வினை வடிவங்களின் குழுமம் பெயர்த் தொடருக்கப் பின்வந்து பின்னுருபுத் தொடராய் அமையும். பின்னுருபாய் வரும் பல பெயர், வினை வடிவங்கள் முறையே தங்கள் பெயர், வினை தொடரியல் இயல்புகளை இழந்து விடுகின்றன. பின்னுருபுகளை அவற்றின் வடிவ மற்றும் அவை எந்தப் பெயர்த் திரிபுக்குப் பிறகு வருகின்றன என்பதில் அடிப்படையில் வனைப்பாடு செய்யலாம். எடுத்துக்காட்டாக, இல் ஒட்டு எடுத்து வரும் நடுவில்,இடையில் என்பன – உக்கு வேற்றுமைக்குப் பின்னர் வரும். (எ.கா. அவர்களுக்கு நடுவில்,இடையில்) வினைவழிப் பின்னுருபுகளை இறந்தகால வினை எச்ச வடிவில் வருவன. (எ.கா. இல்லாமல்,அல்லாமல்) காலங்காட்டா வினையெச்ச வடிவில் வருவன. (எ.கா. தவிர, போல, விட) என வகைப்படுத்தலாம்.

பெயரடைகள்

ஆக்ஷர், லேமன், கோதண்டராமன் ஆகிய மூவருமே பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்கின்றனர். பெயரடையை ஒரு சொல் வகுப்பாகக் கொள்வதில் மொழியியல் அறிஞர்களிடையே ஒருமித்த கருத்து இல்லை. பெயரடைக்கு ஒரு செயல்முறை வரையறை செய்வது எளிதானதல்ல. பெயரடை தொடரியல் அடிப்படையிலான ஒரு சொல் வகையாகும். பெயரடை, பெயர்த்தொடரில் தலைப்பெயருக்கு முன்னதாக அதன் அடையாக வரும் (எ.கா. பெரிய மரம்), வரையறு அடை (Determiner) பெயரடைக்கு முன் வரலாம். (எ.கா. இந்தப் பெரிய வீடு) பயனிலை இடத்தில் பெயரடை வருவதானால் அது மாற்றுப்பெயர் ஒட்டாக்கம் பெற்று வர வேண்டும்.(எ.கா. அந்த மாளிகை பெரியது) சொல் வகுப்புகளில் பெயரடை குறைச்சொல் எடுப்பதில்லை. பெயரடைகளைத் தனிநிலைப் பெயரடை (எ.கா.நல்ல, பெரிய, கெட்ட, சிறிய), கூட்டுநிலைப் பெயரடை மற்றும் ஆக்கப்பெயரடை (derived adjective) (எ.கா. அழகான, உயரமான) என வகைப்படுத்தலாம். தனிநிலைப் பெயரடைகள் மரபிலக்கணத்தில் குறிப்புப் பெயரெச்சங்களாகக் கொள்ளப்படுகின்றன. பெயருக்கு பின் வந்து அவற்றைப் பெயரடைகளாகச் செயலாற்றச் செய்யும் ஆன, உள்ள, அற்ற, இல்லாத (எ.கா.அழகான, அழகற்ற, அழகில்லாத) என்ற பின்னுருபுகளைப் பெயரடையாக்கப் பின்னொட்டுகளாகக் கொள்வதா? அல்லது பெயரெச்சங்களாகக் கொள்வதா? என்பது இலக்கணப் பிரச்சனை. பெயரடைகளும் (தெரிநிலை) பெயரெச்சத் தொடர்களும் செயற்பாடு அடிப்படையில் பெயருக்கு அடையாய் வருவன (எ.கா. அந்த நல்ல பையன், அந்த படித்தப் பையன்) ஆனால் (தெரிநிலை) பெயரெச்சங்கள் உடன், பின், பிறகு, போது, முன், மட்டும், வரைக்கும் முதலான உருபுகளுடன் வினையடைத் தொடர்புகளை உருவாக்கும். (எ.கா. வந்தவுடன், வந்தபின், வரும்முன், கூடியமட்டும்) பெயரடைகள் (குறிப்புப் பெயரெச்சங்கள்) இவ்வாறு செயல்படுவதில்லை. (பரமசிவம் (1983:194)) பரமசிவம் மரபிலக்கணங்கள் பெயரெச்சங்கள் எனக் குறிப்பிடும் தெரிநிலைப் பெயரெச்சம் (எ.கா. பாடின பெண்) குறிப்புப் பெயரெச்சம (எ.கா.நல்ல பெண்), எதிர்மறைப் பெயரெச்சம்

(எ.கா. படிக்காத பெண்) என்ற மூன்றையும் ஆன ஒட்டாக்காப் பெயரடைகளையும் (எ.கா.அழகான பெண.) பெயரடைகளாகக் கொண்டு அவற்றில் தெரிநிலைப் பெயரெச்சத்தையும், எதிர்மறைப் பெயரெச்சத்தையும் தொடர் என்று குறிப்பிட்டு ஒரு வகையாகவும், குறிப்புப் பெயரெச்சத்தையும் ஆன ஒட்டாக்கப் பெயரடையையும் தனிச்சொல்லாகக் கொண்டு மற்றொரு வகையாகவும் பிரிக்கிறார்.

வினையடைகள்

ஆஷர் (Asher, 1982:101-102) 'Operational Definition for Word Classes' என்ற தலைப்பின் கீழ் சொல் வகுப்புகளில் ஒன்றாக வினையடையைத் தராவிடிலும் சொல்லாக்க உருபனியலைப் பற்றிப் பேசுகையில் அவர் (Asher, 1982:199-205) விடையடைகளின் ஆக்கம் பற்றிக் குறிப்பிடுகிறார். வாக்கியத்தில் அவை இடம் பற்றிக் கூறுகையில் இடவாக்கியங்கள் (Locative Sentences) மற்றும் உள்ளமை வாக்கியங்கள் (existential Sentences) அல்லாத பிற வாக்கியங்களில் வினையடை பொதுவாக எழுவாய் மற்றும் அயல் செயப்படுபொருளைத் தொடர்ந்தோ நேர்ச் செயப்படுபொருளுக்கு முன்னரோ வரும் என்று குறிப்பிடுகிறார். ஒரே வாக்கியத்தில் வேறுபட்ட வினையடைகள் வந்தால் அவை வரும் நிரலைச் சரியாகக் கூறமுடியாது. பொதுவாக, கால வினையடைகள், இட வினையடைகளுக்கு முன்னர் வரும். இடவாக்கியங்கள் இவற்றில் வினையடை எழுவாய் முன்னர் வரும். கோதண்டராமனும், லேமனும் வினையடையைச் சொல் வகுப்பாகக் கொள்கின்றனர். வினையடைகளைத் தனிநிலை வினையடை (Simple adverbs) (எ.கா.நன்கு, உரைக்க) கூட்டுநிலை வினையடை அல்லது ஆக்க வினையடை (எ.கா. அழகாக, அழகாய்) எனப்பகுக்கலாம். திரிபுற்ற அல்லது திரிபுறாத பல பெயர்ச்சொல் மற்றும் வினைச்சொல் வடிவங்களை தொடரியல் அடிப்படையில் நெருங்கிய ஒரு குழு வினையடையாக மீள் ஆய்வு செய்யலாம். பின்னுருபுகளாக வரும் பல சொல் வடிவுகள் வினையடைகளாகவும் வரலாம். (எ.கா. முன்னர், முன்னால், பிறகு) திரிபுறாத ஆனால் குறைச்சொல் ஏ –ஐ

விருப்பாக ஏற்றுவரும். பெயர்ச்சொல் வடிவுகள் (எ.கா. அப்பாலே, உள்ளே, அப்புறம்) இடவேற்றுமைக்காகத் திரிபுற்ற பெயர்ச்சொல் வடிவுகள் (எ.கா. பார்த்து, பிந்தி, முந்தி) என்பன வினையடைகளாகவும், பின்னுருபுகளாகவும் செயல்பட வல்லன. பெயரல்லாத வேறுசில சொல்வடிவுகளும் வினையடையாகச் செயல்பட வல்லன. எடுத்துக்காட்டாகச் சுட்டு மற்றும் வினாச் சுட்டுரபுகள் கொண்ட அப்படி, எப்படி, அப்போது, எப்போது, இங்கு, எங்கு, இன்று, என்று என்பனவற்றையும், பொதுவாக வினையடைகளாகக் கொள்ளலாம். சில இடங்களில் இவை வினையடையல்லாது பெயர்போல் செயற்பாடு செய்யலாம். (எ.கா. இன்று நல்ல நாள்) பல திரிபுற்ற மற்றும் திரிபுறாத பெயர் மற்றும் வினை வடிவங்கள் தொடரியல் அடிப்படையில் வாக்கியநிலை வினையடைகள் அல்லது வினையடை இணைப்பான்கள் என மீள் ஆய்வு செய்யப்படலாம். இவை வாக்கியத்தின் முன்னிடத்தில் வந்து இரண்டு வாக்கியங்களின் பொருண்மை உறவைக் கூறி நிற்கின்றன. எ.கா. அவன் நன்றாய்ப் படித்தான், ஆனால் வெற்றிபெறவில்லை. அப்படியும், ஆனால், இருந்தாலும், எதற்கும் என்பன வாக்கியநிலை வினையடைகளாகும்.

பெயர்ச்சொல்லுடன் ஆகு என்ற வினைச் சொல்லின் காலமிலா வினையெச்ச வடிவான ஆக அல்லது ஆய் சேர்ந்து கிடைக்கும் சொல் வடிவங்களை ஆக்ஷரும், கோதண்டராமனும் வினையடைகளாகக் கொள்கின்றனர். ஆனால் லேமன் ஆக,ஆய் என்பது வினையடையாக்கம் செய்வது மட்டுமன்றி பிற செயற்பாடுகளிலும் வருவதைச் சுட்டிக்காட்டி ஆக,ஆய் என்பதை கட்டுநிலைப் பின்னுருபு அல்லது குறைச்சொல் என்று ஆயலாம் என்கிறார்.

பொருண்மை அடிப்படையில் வினையடைகளைப் பின்வருமாறு பகுக்கலாம்:

1. கால வினையடைகள் (எ.கா.இன்று, நாளை)
2. இட வினையடைகள் (எ.கா.இங்கு, அங்கு)

3. முறைமை வினையடைகள் (எ.கா.மெல்ல, நன்கு)

4. செயல்தொடர் வினையடைகள் (எ.கா.அடிக்கடி, மறுபடியும்)

5. அளவு வினையடைகள்

ஆனால் லேமனும், ஆஷ்ரும் அளவு வினையடைகளை ஒரு தனி வகுப்பாகக் கொள்கின்றனர். பரமசிவம்(1983) வினையெச்சங்களையும் ஒருவகை வனையடையாகக் கொள்கின்றார். வினையெச்சங்கள் காலக்குறிப்பினைக் கொண்டு விளங்குவதால் இவற்றை வினையடைகளாக எல்லோரும் கொள்வதில்லை.

அளவயடைகள்

ஆஷரும், லேமனும் அளவயடைகளைச் சொல் வகுப்பாகக் கொள்கின்றனர். அளவடைகளுக்கு உருபனியல் அடிப்படையில் ஒட்டு மொத்தமான வரையறை தர இயலாது. நெருங்கிய ஒரு குழுமச் சொல்வடிவுகள் (எ.கா. சற்று, முழு, கொஞ்சம், இத்தனை, அத்தனை, இவ்வளவு, அவ்வளவு, எவ்வளவு, நிரம்ப, நிறைய, மிகவும்) பெயர் அடையாக வர வல்லன. ஆனால் இவை வருமிடங்கள் ஒன்றானதல்ல. முழு என்ற அளவடை எப்பொழுதும் தலைப்புப் பெயருக்கு முன் அதனை அடுத்து வரும். (எ.கா.சற்று நேரம்), பெயருடன் ஆக,ஆன, இல்லாத, உடைய சேர்ந்த சொல் வடிவங்களின் முன்னர் மிகவும் வரும். (எ.கா. மிகவும் அழகாக, மிகவும் அழகான, மிகவும் அழகில்லாத) பிற அளவடைகள் பெயரடைரூபெயர் என்ற தொடரின் முன்வரும். (எ.கா.கொஞ்சம்,பெரிய,பாத்திரம்) முழு, இத்தனை, அத்தனை, எத்தனை, மிகவும், தவிர்த்த பிற அளவடைகள் வினையின் அடையாக வினை முன்னிடத்தில் வரும். எ.கா. சற்று மறந்தான், நிறைய சாப்பிட்டான்). மிகவும் என்பன பெயரடையின் அடையாக வந்து பெயரடைத் தொடரில் காணப்படும். (எ.கா கொஞ்சம் சின்னக் கை) கோதண்டராமன் மிகவும் என்பதை பெயர், வினை,

பெயரடை, வினையடை ஆகிய நான்கினுக்கு முன் அடையாக வரும். சொற்களை வல்லடை என்ற தனிச்சொல் வகுப்பில் அடக்குகின்றார். (எ.கா. மிகவும் கறுப்பு, மிகவும் நல்ல, மிகவும் வேகமாக, மிகவும் பிடிக்கும்).

வரையறை அடைகள்

பெயருக்குப் பின்னர் வரும் இந்த, அந்த என்ற சுட்டுப்பெயரடைகள் சரையறை அடைகளாகும். ஆவை சுட்டப்படும் பொருள் அல்லது குறிப்புப்பொருள் (referent) பேசுபவர்களிடமிருந்து எவ்வளவு அண்மைப்பட்டு அல்லது சேய்மைபட்டு இருக்கின்றது என்பதைக் குறிக்கவல்லன. (எ.கா இந்தப் பையன், அந்தப் பெண்) கிரியாவின் தற்காலத் தமிழ் அகராதியில் இவை சுட்டுப் பெயரடைகள் என்று குறிப்பிடப்படுள்ளன.

இணைப்பான்கள்

கோதண்டராமனும், லேமனும் இணைப்பான்களைச் சொல்வகுப்பாகக் கொள்கின்றனர். "இரண்டு சொற்றொடர்களை அல்லது சொற்களை இணைப்பதாகவும் ஒரு முழுத்தொடரைத் தன்னோடு இணைப்பதாகவும் அமைந்த சொற்கள் இணைப்பான்கள்" எனக் கோதண்டராமன் விளக்குகின்றார். தமிழில் பொதுவாக குறைச்சொற்கள் (clitics) இருப்பினும் பலவினை வடிவங்களை தொடரியல் அடிப்படையில் இணைப்புச சொற்களாகக் கொள்ளலாம். (எ.கா. ஆனால், அல்லது, இல்லையென்றால்) இணைப்புச் சொற்கள் இருபெயரடைகளையோ, பெயர்களையோ இணைக்கப் பயன்படும். (எ.கா. பழைய ஆனால் நல்ல புத்தகம், ரேசஷ்மா இல்லையென்றால் கலா வருவாள். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இணைப்பு இடைச்சொற்கள் என்று குறிப்பிடுகின்றது.

குறைச்சொற்கள்

குறைச்சொற்கள் பற்றி ஆரோக்கியநாதன் (1982) விரிவாக ஆராய்ந்துள்ளார். கோதண்டராமன் குறைச்சொற்களைத துணைமைச்சொல் வகுப்பாகச் சார்புநிலைக்கிளவி வகையின் கீழ் **'ஒட்டு'** என்று குறிப்பிட்டு வகைப்படுத்துகிறார். அவர் "தொடரில் பல்வேறு இடங்களில் நின்று தொடர்ப் பொருளில் நிலைமாற்றம் செய்யும் மற்றும் பெயர் விகுதியாகவோ, வினைவிகுதியாகவோ" கொள்ள இயலாத தான், உம், ஆ முதலான கிளவிகளை ஒட்டுகள் என்று குறிப்பிடுகின்றார். லேமன் குறைச் சொற்களைப் பற்றிப் பேசும் போது உருபனியல் அடிப்படையில் அல்லாமல் இலக்கணத்தின் ஒலியனியல் விதிகளின்படி சொறகளுள் ஒட்டப்பெறும் கட்டு வடிவங்கள் குறைச்சொற்கள் எனக் குறிப்பிடுகின்றனர். அவை திரிபு அல்லது ஆக்கக் கூறுகளின் பிரதிபலிப்பல்ல. திரிபு அல்லது ஆக்கப் பின்னொட்டுகளைப் போல் அவை ஒரு குறிப்பிட்ட சொல் வகுப்பை மட்டும் சார்ந்து வருவன அல்ல. தமிழில் எல்லாக் குறைச்சொற்களும் பின் குறைச் சொற்களாகும். பெயரடை அல்லது பெயரின் அடையாய் செயல்படும் பெயர்கள் தவிர பிற தொடரியல் வகைப்பாடுகளின் தலைமைச் சொற்களுடன் குறைச் சொற்கள் ஒட்டப்பெறலாம். ஒரு குறிப்பிட்ட ஒலியனியல் வடிவம் கொண்ட குறைச்சொல்லுக்குப் பல பொருண்மைச் செயல்பாடுகள் இருக்கும். உம், ஓ, ஏ, தான், ஆ என்பன குறைச்சொல் வகுப்பின் பாற்படும். ஆக்ஷர் ஏ, தான் ஆகிய தேற்றச் குறிகளையும் (emphatic markers) ஆக என்ற வினாக் குறியையும் (interrogative marker) உம், ஓ ஆகிய இணைப்பான்களையும் பலதரப்பட்ட முக்கியமான வாக்கிய உறுப்புகளுடன் ஒட்டத் தகுந்த மாற்றம் கொள்ளாத கட்டு வடிவங்கள் எனக் குறிப்பிட்டு அவற்றை இடைச்சொற்கள் (Particles) என்ற சொல் வகுப்பில் அடக்குகின்றார். கிரியாவின் தற்காலத் தமிழகராதி இவற்றை இடைச்சொற்கள் எனக் குறிப்பிடுகின்றது.

வினைசார் கிளவி

கோதண்டராமன் வந்தபின், வந்தபோது, வந்த உடன் ஆகியவற்றில் காணும் பின், போது, உடன் ஆகிய கிளவிகளை வினைசார் கிளவி என துணைச்சொல் வகையாகச் சார்புநிலைக் கிளவி வகையின் கீழ் தருகின்றார். "பெயர்த்தன்மை திரிந்து பெயரெச்சத்தின் பின் விகுதி போல நின்று வினை எச்சங்களை அமைக்கும் கிளவிகளும், வினைசார் கிளவிகள்" என்று குறிப்பிடுகின்றார். செய்யத்தக்க, செய்யக்கூடிய, செய்ய வேண்டிய முதலான கூட்டுநிலைப் பெயரெச்சஆகளில் காணும் தக்க, கூடிய, வேண்டிய என்பனவற்றையும் வினைசார் கிளவிகளாகக் கொள்ளத்தகும் என்கிறார். கிரியாவின் தற்காலத் தமிழகராதி இயற்றையும் இடைச்சொல் என்றே குறிப்பிடுகின்றது. வேற்றுமைப் பின்னொட்டுகளையும் (எ.கா. ஐ, ஆல், கு) அன், ஆன், அள், ஆள் முதலான பால் இட ஒட்டுகளையும் ஆன, ஆக என வரும் பெயரடை, வினையடை விகுதிகளையும் இந்த வகுப்பில் அடக்குகின்றார்.

சாரியை அல்லது நிரப்பி

"தமக்கென்று அகராதிப் பொருளோ, இலக்கணப் பொருளோ இல்லாதனவாய் சொற்கள் ஒன்றோடொன்று சேரும்போது சொற்சேர்க்கைக்கு இசைவாக நின்று சொல் அமைப்புக்கு உதவுவனவாய் அமைந்த கிளவிகள் சாரியை அல்லது நிரப்பி என்று பெயர்பெறும்" என்று கோதண்டராமன் இச்சொல் வகுப்பை வரையறை விளக்கம் செய்கிறார். வீட்டடினை, ஊரினை என்பனவற்றிலுள்ள இனை என்பதையும், வந்தனன், சென்றனன என்பனவற்றிலுள்ள அன் என்பதையும் சாரியை கிளவிக்கு எடுத்துக்காட்டுகளாகக் காட்டுகின்றார்.

மேலே கூறப்பட்ட அனைத்துச் சொல் வகைப்பாடு தமிழுக்குத் தேவையான அடிப்படைச் சொல் வகைப்பாடுகளாகும். இதை தவிர பல நுண்மையான வகைப்பாடுகளையும் உட்படுத்தலாம். எடுத்துக்காட்டாக, வினையை, வினையெச்சம், பெயரெச்சம், வினைமுற்று, துணைவினை என மேலும்

வகைப்படுத்தலாம். இச்சொல் வகைப்பாடுகள் சொல்வகை அடையாளக் குழுமத் தேர்வுக்கு அடிப்படையாய் அமையும்.

4.6.2 தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

4.6.2.1 CIIL நிறுவனத்தின் சொல் வகைப்பாட்டு அடையாளக் குழுமங்கள்

இந்திய மொழிகளுக்கு தரவுத்தொகுதிகள் உருவாக்குவதில் முன்னோடியாக விளங்கிய இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்திய மொழிகளுக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமங்களை உருவாக்கித் தரவுத்தொகுதிகளை அடையாளப் படுத்தியுள்ளது. 2006-இல் இந்தி தரவுத்தொகுதிக்கு சொல் வகைப்பாட்டு அடையாளக் குழுமமாக முடிவு செய்யப்பட்ட சொல் வகைப்பாட்டு குழுமத்தின் பட்டியல் கீழே தரப்பட்டுள்ளது.

CIIL Hindi Corpora Tagset Finalized in August 29.2006

NOUN

1. NN	Commom Noun
2. NNP	Proper Noun (Palce, Person,etc.…)
3. NC	Noun Compound
4. NAB	Noun Abstgract
5. CRD	Cardinal number
6. ORD	Ordinal number

PRONOUN

7. PRP	Personal Pronoun
8. PRI	Indefinite Pronoun
9. PRR	Pronoun Reflexive
10. PRL	Relative Pronoun
11. PDP	Demonstrative Pronoun

VERB

12. VF Verb Finite Main
13. VNF Verb non-finite Adjectival/ Verb non-finite Adverbial
14. VAX Verb Auxiliary
15. VNN Gerund/Verb non-finite nominal
16. VINF Verb Infinitive
17. VCC Verb Causative
18. VCD Verb Double Causative

ADJECTIVE

19. ADD Adjective - declinable
20. ADI Adjective - Indeclinable

ADVERB

21. ADV Adverb
22. ADL Adverb of Location

OTHERS

23. QWO Question Words
24. PPO Post position
25. INT Intensifier
26. QTF Quantifier
27. CNJ Conjunction
28. INJ Interjection
27. IND Indiclinable
30. QOT Quotative
31. PRT Particle
32. NEG Negative
33. RDP Reduplication
34. FWD Loan Word [non – nativized words]
35. IDM Idiom

36. PRO Proverb

4.6.2.2 AU-KBC நிறுவனத்தின் தமிழுக்கான சொல் வகைப்பாட்டு அடையாளக் குழுமம்

AU-KBC நிறுவனம் தமிழுக்கான ஒரு சொல் வகைப்பாட்டுக் குழுமத்தை உருவாக்கியுள்ளது. இது எல்லா இலக்கண மற்றும் சொல்சார் உறுப்புகளையும் உள்ளடக்கிய குழுமம் ஆகும். இது தமிழ்ப்பல்கலைக் கழக மொழியியல் வல்லுநர்கள் உதவியால் AU-KBC -இன் NLP குழுவினரால் 2001-இல் திட்டமிடப்பட்டு உருவாக்கப்பட்டதாகும். பின்வரும் பட்டியல் தமிழின் சொல் வகைப்பாட்டு அடையாளக் குழுமமாகும்.

TAMIL TAGSET

No	Description	TAG	Examples	Suffixes
1	Noun Singular	N.Sg	மரம் (tree)	
2	Noun Plural	N.pl	மரங்கள் (trees) பூக்கள் (flowers)	
3	Pronoun	Pr.Sg	avanY (he), nAnY(I)	
4	Pronoun Oblique from	Pr.Obl.Sg	என்(mine)	
5	Pronoun plural	Pro.Pl	அவர்கள் (they)	
6	Pronoun Plural Oblique	Pro.Obl.Pl	எங்கள் (ours), உங்கள் (yours)	
7	#Euphonic Increment	Eu	பசுவின் (of the cow)	
8	Acusative	Acc	அவனை (him	ஐ

			accusative)	
9	Dative	Dat	அவனுக்கு (to)	உக்கு, கு, க்கு
10	Instumental	Inst	சாவியால் (by the key)	ஆல்
11	Sociative	Soc	அவனோடு (with him), அவனுடன் (with him)	ஓடு, உடன்
12	Locative	Loc	கையில் (in the story)	இல்
13	Ablative	Abl	ஊரிலிருந்து (from city/town	ல்லிருந்து
14	Benefactive	Ben	அவளுக்காக (for her)	உக்காக
15	Genetive	Gen	பசுவின் (of the cow), உடைய (of), அவனது (his)	இன், உடைய, அது
16	Vocative	Voc		அட, இவளே
17	Clitics	Cl		ஆ,ஏ,ஓ,உம்,தான்
18	Verb	V	ஓடு (run)	
19	#Transitive Verb	VT	திருப்பு (turn sth)	
20	#Intransitive Verb	VI	திரும்பு (turn)	
21	Causative	VC	தெரிவி (inform)	
22	Infinitive	Inf	செய்ய (todo)	அ
23	Singular Negative [imperative]	Neg	செய்யாதே (you don't do)	ஆதே

24	Plural Imperative	V.pl	செய்யுங்கள் (you do-honor plu)	உங்கள்
25	Plural Negative [imperative]	Pl..Neg	செய்யாதிருங்கள் (you do-hon or plu)	ஆதிருங்கள்
26	Imperative+Suffixes	Imp	வாட, அம்மா	
27	Present Tense	Pre		கிற், க்கிற், கின்ற், க்கின்ற்
28	Past Tense	Pst		த், ந்த், இன்,த்த், இ, ட், ற்
29	Future Tense	Fut		வ்,ப், ப்ப், உம்
30	First person Singular	1s		என்
31	First person Plural	1pl		ஓம்
32	Second person Singular	2s		ஆய்
33	Second person Plural	2pl		ஈர், ஈர்கள்
34	Third person masculine Singular	3Ms		ஆண்
35	Third person feminine Singular	3 Fs		ஆள்
36	Third person Honorific Singular	3 Hs		ஆர், அர், அனர்
37	Third person Honorific Plural	3 H pl		ஆர்கள்
38	Third person singular	3 N S		awu, wu, um

	Neuter (non-human)			
39	Third person plural Neuter (non-human)	3 Npl		அன, உம்
40	Third person Neuter Future Negative	3NFut Neg	விடாது (it/they), (won't)	ஆ (ஆது?)
41	Future Negative human	Hum Neg	மாட்டேன் (I don't)	மாட்ட்
42	Present/Past Negative (non-future)	Nfut Neg	வரவில்லை (didn't come)	இல்லை
43	#Optative	Opt	வாழ்க (long live)	க, க்க
44	Verbal Participle	Vbp	செய்து (done)	து, ந்து, etc.,
45	Verbal Participle Negative	Vbp Neg	செய்யாமல் (non-done)	ஆமல், ஆது
46	Contitional Positive	Cnd	செய்தால் (if done)	ஆல்
47	Conditional Negative	Cnd Neg	செய்யாவிட்டால் (if not done)	ஆவிட்டால்
48	Adjectval Participle Present/Past	AdjP	செய்த (do pre/pst+sth), செய்கிற (do pre/Pst-th)	த, கிற
49	Adjectival Participle Future	Adjp Fut	ceVyyum (do Future sth)	Um
50	Adjectival Participle Negative	Adjp Neg	செய்யாத (don't do sth)	ஆத
51	Verbal Noun Untensed	Vbn Unt	செய்யல் (doing), செய்தல் (doing), படிக்கல் (reading),	அல், தல், த்தல், க்கல், கை, க்கை

			படித்தல்l (reading), செய்கை (doing), படிக்கை (reading)	
52	Verbal Noun Tensed Positive	Vbn	செய்தது (have done)	ஆது
53	Verbal Noun Tensed Negative	Vbn Neg	செய்யாத்து (have not done)	ஆதது
54	Participal Noun	PpIN	செய்தவன் (one who did)	அன், அள்,து, அர்,அர்கள்.ஐ
55	Adjctival Noun	AdjN	நல்லவன் (good man)	அன், அள்,து, அர், அர்கள்.ஐ
56	Selective (Concessive?)	Sel	அவன், ஆவது(at least him)	ஆவது
57	Ordinal Suffix	Ord	முதலாவது(First)	ஆவது
58	Reportative Suffix	Rep	-	ஆம்
59	Complementizer	Cmp	-	என்று, என்பது, ஆக
60	Auxiliary	Aux	-	-
61	Conjunction	Conj	-	ஆனால், இல்லை என்றால், அல்லது
62	Adjective	Adj	அழகான(beautiful with adj suffix)	ஆன
63	Adverbs	Adv	அழகாக(beautiful with adv Suffix)	ஆக
64	Postposition	PSP	மேல் (above), கீழ்	

			(below)	
65	Quantifiers	Q	ஒன்று (one), சில (few), கொஞ்சம் (some),	
66	Determiners	Det	அந்த (that), இந்த (this)	
67	Wh. Words	INT	என்ன(what), எவை(which), எப்படி(how)	
68	Unknown	UNK	-	-

4.6.2.3 அமிர்தா தொழில்நுட்பக் கல்லூரி பயன்படுத்திய தமிழுக்கானச் சொல்வகை அடையாளக் குழுமம்

அமிர்தா தொழில் நுட்ப கல்வி நிறுவனத்தில் தமிழை மொழிபெயர்ப்பதற்காக நுட்பங்களைப் பயன்படுத்தி அமிர்தா தொழில் நுட்ப கல்லூரி பேராசிரியர்களான தனலெட்சுமி, அனந்த் குமார், சோமன் மற்றும் தமிழ்பல்கலைக்கழக மொழியியல் துறைத்தலைவர் ச. இராஜேந்திரன் போன்றோர்களின் கூட்டு முயற்சியால் உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்திகளின் அட்டவணையைக் கீழேக் காணலாம்.

S.No	POS	DESCRIPTION
1	NN	NOUN
2	NNC	COMPOUND NOUN
3	NNP	PROPER NOUN
4	NNPC	COMPOUND PROPER NOUN

5	ORD	ORDINALS
6	CRD	CARDINALS
7	PRP	PRONOUN
8	PRIN	PRONOUNINTROGATIVE
9	PRID	PRONOUN INDIFINITE
10	ADJ	ADJECTIVES
11	ADV	ADVERB
12	VNAJ	VERB NON-FINITE ADJECTIVES
13	VNAV	VERB NON-FINITE ADVERBS
14	VBG	VERBIAL GERLND
15	VF	VERB FINITE
16	VAX	VERB AUXILARY
17	VNIT	VERB IN FINITE
18	CNJ	CONJUNTION
19	CVB	CONDITIONAL VERB
20	QW	QUESTION WORD
21	COM	COMPLEMENTIZER
22	NNQ	QUANTITY NOUN
23	QTF	QUANTIFIERS
24	PPO	POSTPOSITION
25	DET	DETGERMINERS
26	INT	INTENSIFIER
27	ECH	ECHO WORDS
28	EMP	EMPHASTS

29	COMM	COMMA
30	DOT	DOT
31	QM	QUESTION MARKS
32	RDW	REDUPLICATION

இவ்வட்டவணையைக் கொண்டு உருவாக்கப்பட்ட சொல்வகை அடையாளப்படுத்தியைக் கொண்டு பல்வேறு அடிப்படையில் உருவாக்கப்பட்ட தொடர்களின் அமைப்பைக் கீழேக் காணலாம்.

பெயர்த்தொடர்

[அந்த <DET> (B-NP) அழகான <ADJ> (I-NP) பறவை <NN> (I-NP)] NP

பெயரடைத்தொடர்

[திரைப்படம் <NN> (B-AJP) சார்ந்த <ADJ> (I-AJP)] AJP

வினையடைத்தொடர்

[அருகே <ADV> (B-AVP)]AVP

இணைப்புச்சொல்

[ஆனால் <CNJ>(B-CJP)] CJP

[என்று<COM> (B-COMP)] COMP

வினைமுற்று தொடர்

[உள்ளது <VF> (B-VFP)] VFP

வினைமுற்றுப்பெறாத வாக்கியம்

[வெளி வந்த (VNAJ) (B-VNP)] VNP செய்திக் <NNC> < B-NP> குறிப்பு <I – NP> <NNC>

[விரைந் து <VNAV>(B-VNP)] VNP செய்தான் <VF>

gerundial chunk

தொழிற்சாலை <NN> [அமைப்பதில் <VBG>(B-VGP)] VGP தாமதம் <NN>

ஒரு விரிதரவில் எவ்வாறு சொற்களை சொல்வகை அடையாளப்படுத்தலாம் என்பதற்கான ஒரு எடுத்துக்காட்டைப் பார்ப்போம்.

வளாகத் <NNC> <B-NP>

தேர்வில் <NNC> <I-NP>

வேலைவாய்ப்பு <NN> <B-NP>

பெற்ற <VNAJ> <B-VNP>

மாணவர்களின் <NN> <B-NP>

பட்டியல் <NN> <I-NP>

வெளியிடும் <VNAJ> <B-VNP>

விழா <NN> <B-NP>

திங்கள்கிழமை <NNP> <B-NP>

நடைப்பெற்றது<VF> <B-VFP>

. <DOT> <O>

4.4.2.4 IIIT ஹைதராபாத்தில் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக் குழுமம்

இந்திய மொழிகளுக்கிடையிலான மொழிபெயர்ப்பிலும், ஆங்கில மொழியிலிருந்து இந்திய இயந்திர மொழிப்பெயர்ப்பிலும் சூடுப்பட்டுள்ள ஷைதராபாத்தில் உள்ள IIIT நிறுவனம் இந்தியாவில் இயற்கை மொழி ஆய்வுக்கு முன்னோடியாக உள்ளது. இந்நிறுவனம் உருவாக்கிய இந்திய மொழிகளுக்கான (குறிப்பாக இந்திக்கான) சொல்வகைப்பாட்டு அடையாளக் குழுமம் பரவலாகப் பயன்படுத்தப்படுகின்றது. இதன் பட்டடியல் கீழே தரப்பட்டுள்ளது.

NN	Common Noun
NNP	Proper Noun
*C	for all Compounds
QC	Cardinal No

QO Ordinal No

PRP Pronoun

VF Verb finite main

VNF Verb Non-finite adverb

VAUX Verb Auxiliary

VNN Gerund/Verb non-finite nominal

VINF Verb Infinitive

JJ Adjectives

RB Adverbs

PSP Post position

QF Quantifiers

WQ Question words

CC Conjunction

NEG Negative

INT Interjection

CL Classifier

SVM Special

4.6.2.5 AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு

AU-KBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீட்டுப் பட்டியல்கள் கீழே தரப்பட்டுள்ளது. இப்பட்டியல் நிறுவன அறிவியலாளர் எல். சோபா என்பவரால் உருவாக்கப்பட்டது.

NOUNS

AU-KBC	CIIL	IIT-H
	NN - Common Noun	NN - Common Noun

	NNP - Proper Noun	NNP - Proper Noun
	NC - Noun Compound	*C - for all Compound
	CRD - Cardinal No	QC - Cardinal No
Ord - Ordinal No	ORD - Ordinal No	QO - Ordinal No
N.Sg - Noun Sg		
N.Pl - Noun Pl		
	PRP – Personal pronoun	PRP –Pronoun
	PRI – Idenfinite Pronoun	
	PRR – Reflexive Pronoun	
Pr.sg – Pronoun Sg		
Pr.Pl – Pronoun Pl		
Pr.Obl.Sg – Pronoun Oblique Sg		
V – Verb	VF – Verb Finite Main	VF - Verb Finite Main
	VNF – Ver non finite Adverbial adjectival	VNF – Ver non finite Adverbial adjectival
Vbn – verbal noun untensed, Vbn – Vbn Tensed Positive, Vbn Neg – Vbn Tensed Negative	VNN – Gerund/Verb non-finite nominal	VNN – Gerund/Verb non-finite nominal
Inf – Infinitive	VINF – Verb Infinitive	VINF – Verb Infinitive
VC – Verb causative	VCC – Verb Causative	
	VCD – Verb Double Cousative	

VT – Verb Transitive		
VI – Verb Intransitive		

ADJECTIVES

AU-KBC	CIIL	IIIT-H
Adj – Adjective		JJ – Adjective
	ADD – Adjective Declinable ADI – Adjective Indeclinable	

ADVERBS

AU-KBC	CIIL	IIIT-H
Adv – Adverb	ADV – Adverb ADL – Adverb of Location	RB – Adverb

OTHERS

AU-KBC	CIIL	IIIT-H
PSP – Postposition	PPO –Post position	PSP – Post position
Q – Quantifiers	QTF – Quantifiers	QF – Quantifiers
CNJ – Conjunction	CNJ - Conjunction	CC – Conjunction
	INT – Intensifer	INT – Intensifer
	NEG – Negative	NEG – Negative
	INJ - Interjunction	INJ – Interjunction
	PRT – Particle	RP – Particle

OTHERS

CIIL
IND – Indeclinable
QOT – Quotative
RDP – Reduplication
LWD – Loan Word
IDM – Idion
PRO – Proverb

இங்கு சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் பற்றிய முந்தய முயற்சிகள், சொல்வகைப்பாடு அடையாளப்படுத்தலுக்கு ஒரு தரமான அமைப்பாக விளங்குகிறது. அவை CIIL சொல்வகைப்பாட்டு அடையாளக் குழுமங்கள் AUKBC நிறுவனத்தின் உருவாக்கப்பட்ட இந்திய மொழிகளுக்கான சொல்வகைப்பாட்டு அடையாளக்குழுமம் AUKBC, CIIL, IIIT-H நிறுவனங்களின் சொல்வகைப்பாட்டு அடையாளக் குழுமங்களின் ஒப்பீடு குறிக்கப்பட்டுள்ளது. இவைகளை தேவை அடிப்படையில் பயன்படுத்த தேர்ந்தெடுத்துக் கொள்ளலாம்.

4.6.2.6 பிற சொல்வகை அடையாளப் படுத்திகள்

தமிழுக்கு வேண்டி பல சொல்வகை அடையாளப்படுத்திகள் உருவாக்கப்பட்டுள்ளன. அவற்றில் சில முக்கியமானவை கீழே தரப்பட்டுள்ளன.

4.6.2.6.1 வாசு அரங்கநாதனின் டேக் தமிழ்

வாசு அரங்கநாதனால் உருவாக்கப்பட்ட டேக் தமிழ்ச் (Vasu Ranganathan's Tag Tamil) சொல்-ஒலியனியல் (Lexical-phonology) அணுகுமுறை அடிப்படையில் உருவாக்கப்பட்டது. டேக் தமிழ் வினைகளின் உருபனியல் பகுப்பாய்வை உருபன்களின் வருகை முறையை வரிசை அட்டவணையைப் பயன்படுத்தி செய்கினறது. டேக் தமிழ் சொல் வகை அடையாளப்படுத்தலையும் உருவாக்கலையும் செய்கிறது.

4.6.2.6.2 கணேசனின் சொல்வகை அடையாளப்படுத்தி

கணேசன் தமிழுக்கு வேண்டி ஒரு சொல் வகை அடையாளப்படுத்தியை உருவாக்கியுள்ளார். அவருடைய அடையாளப்படுத்தி இந்திய மொழிகளின் மைய நிறுவன தமிழ் விரிதரவில் நன்கு வேலை செய்கின்றது. பிற தரவுத்தொகுதிகளின் இதன் திறன் மதிப்பீடு செய்யப்பட வேண்டும். இவர் தமிழுக்கு வேண்டி விரிவான சொல்வகை அடையாளப்படுத்திகளின் குழுமத்தை வைத்திருக்கின்றார். இவர் இந்திய மொழிகளின் மைய நிறுவன தரவுத்தொகுதியின் ஒரு பகுதியை அகராதியாலும் உருபனியல் பகுப்பாய்வியாலும் சொல் வகைக்காக அடையாளப்படுத்தி உள்ளார். பின்னர் இவை அடையாளப்படுத்தப்பட்ட விரிதரவை மனிதசக்தியால் திருத்தி மீதியுள்ள பெருந்தரவுக்கு பயிற்சி தந்துள்ளார். இவருடைய சொல்வகை அடையாளப்படுத்தியில் சொல்வகை அடையாளப்படுத்தல் ஒவ்வொரு உருபன்களுக்கும் தரப்படுகின்றது.

4.6.2.6.3 RCILTS-இன் தமிழ் கதம்பம்

கதம்பம் ஆவணத்தில் தரப்பட்டுள்ள சொல்லுக்குச் சொல்வகை அடையாளத்தைத் தருகிறது. இது தமிழ்மொழியியல் அடிப்படையிலான ஊக விதிகளைப் பயன்படுத்துகின்றது. அகராதியையோ உருபனியல் பகுப்பாய்வியையோ பயன்படுத்துவதில்லை. இது பெரிய ஆவணங்களில் 80% திறனைத் தருகிறது. இது பன்னிரண்டு ஊக விதிகளைப் பயன்படுத்துகிறது. இது

இட-எண்-பால், காலம் மற்றும் வேற்றுமை குறியீடுகளின் அடிப்படையில் சொல்வகை அடையாளங்களை அடையாளம் காண்கிறது. துனியாக நிற்கும் சொற்கள் சொல்வகை அடையாளப்படுத்தியில் சேகரிக்கப்பட்ட பட்டியலில் இருந்து பரிசோதிக்கப்படுகிறது. இது அறியப்படாத சொற்களுக்கு (unknown words) நிரப்பு விதியைப் பயன்படுத்துகிறது. இது bபைசயஅ மாதிரிகளைப் பயன்படுத்தி முன்வரும் சொல்லின் சொல்வகைப்பாட்டைக் கொண்டு அறியப்படாத சொல்லை அடையாளம் காண்கிறது.

4.7. இந்தி மொழியிலிருந்து இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையில் சொல்வகைப்பாடு அடையாளப்படுத்தல்

தமிழில் முக்கியமான சொல்வகைப்பாடுகள் பெயர், வினை, பெயர்த்தொடர், வினைத்தொடர் என்பவைகளாகும். இந்தியமொழியிலிருந்து இந்தியமொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையில் (இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறை) 21 சொல்வகைப்பாடுக் குழுமங்கள் பயன்படுத்தப்படுகின்றன. சொல்வகைப்பாட்டு அடையாளப்படுத்தி சொல்வகைப்பாட்டு அடையாளப்படுத்தும் இயந்திரத்தால் ஆதரவளிக்கப்படுகின்றது.

4.7.1. இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறையின் சொல்வைப்பாட்டுக் குழுமம்

இ-இ மொழிபெயர்ப்பு ஒழுங்குமுறையில் உட்படுத்தப்பட்டுள்ள எல்லா மொழிகளுக்கும் ஒரு பொதுவான சொல்வகைப்பாட்டு அடையாளக் குழுமம் பயன்படுத்தப்படுகின்றது. இச்சொல்வகைப்பாடுக் குழுமம் தான் தமிழுக்கும் பயன்படுகின்றது. சொல்வகைப்பட்டு அடையாளக் குழுமத்தில் பயன்படுத்தப்பட்டுள்ள அடையாளங்கள் கீளே விளக்கப்பட்டுள்ளன.

1. NN (Noun, பெயர்)

NN என்ற அடையாளம் இலக்கணத் தகவல்கள் அடிப்படையில் வேறுபடுத்தப்படாமல் பொதுவான பெயர்களுக்குப் பயன்படுத்தப்படுகின்றது.

எ.கா. நல்ல <JJ> கோவில் <NN>

2. NST (Noun denoting spial and temporal expressions, இட மற்றும் காலத்தைச் சார்ந்த பெயர்கள்)

முன்னால், *பின்னால்*, *அப்புறம்* போன்ற இட மற்றும் காலம் சார்ந்த சொற்கள் NST எனஅடையாளப் பயன்படுத்தப்படும்.

கோவிலின் <NN>முன்னால் <NST> நிற்கிற ஆள்

3. NNP (Pronoun, இயற்பெயர்)

எல்லா இயற்பெயர்களும் NNP என அடையாளப்படுத்தப்படும்.

4. PRP (Pronoun, மாற்றுப்பெயர்)

எல்லா மாற்றுப்பெயர்களும் PRP என அடையாளப்படுத்தப்படும்.

சென்னையிலிருந்து <PRP> மகாபலிபுரத்திற்கு <PRP> இருபது <QC> கிலோமீட்டர்கள் <NN>.

5. DEM (Demenostrative, சுட்டு)

எல்லாச் சுட்டுக்களும் DEM என அடையாளப்படுத்தப்படும்.

அந்தக் <DEM> கோவில் <NN>

6. VM (Verb main, முதன்மை வினை)

எல்லா முதன்மை வினையளும் VM என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சுற்றுலா <NN> சென்றனர்< VM. <SYM>

7. VAUX (Auxiliary Verb, துணைவினை)

எல்லாத் துணைவினைகளும் VAUX என அடையாளப்படுத்தப்படும்.

அவர்கள் <PRP> சென்னையிலிருந்து <NNP> மகாபலிபுரம் <NNP> செல்ல <VM> புறப்பட்டுக் <VM> கொண்டிருந்தனர் <VAUX>. <SYM>

8. JJ (Adjecive, பெயரடை)

பெரிய <JJ> கோவில் <NN>

9. RB (Adverb, வினையடை)

எல்லா முறைமை வினையடைகளும் RB என அடையாளப்படுத்தப்படும்.

பஸ் <NN> வேகமாகச் <RB> சென்று <MV>கொண்டிருந்தது <VAUX>. <SYM>

10. PSP (Postposition, பின்னுருபு)

எல்லா பின்னுருபுகளும் PSP என அடையாளப்படுத்தப்படும்.

அவர்களுடன் <PRP> கூடவே <PSP> வழிகாட்டியும் <NN> சென்றார் <VM>. <SYM>

11. CC (Conjucts, both co-ordinative and subordinative, இணைப்புகள், சமநிலை இணைப்பும் துணைநிலை இணப்பும்)

சமநிலை இணைப்புகளும் துணைநிலை இணைப்புகளும் CC என அடையாளப்படுத்தப்படும்.

கோயில்கள் <NN> மற்றும் <NN> சர்ச்சுகள் <NN> வழியில் <NN> இருந்தன <VM>.

12. WQ (Question Words, கேள்விச் சொற்கள்)

என்ன, எங்கே, எப்பொழுது போன்ற கேள்விச் சொற்கள் WQ என அடையாளப்படுத்தப்படும்.

13. QF (Quatifiers)

எல்லா அளவை அடைகளும் QF என அடையாளப்படுத்தப்படும்.

பல <QF> கோவில்களைக் <NN> கண்டு <VM>மகிழ்ந்தனர் <VM>.

14. QC (Cardinal, அடிப்படை எண்)

எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

ஒன்று, இரண்டு, பத்து, நூறு போன்ற எல்லா அடிப்படை எண்களும் QC என அடையாளப்படுத்தப்படும்.

இரண்டு <QC> கிலோமீட்டர் <NN> தூரத்தில் <NN>

15. QO (Codinal, முறைமை எண்)

எல்லா முறைமை எண்களும் QO என அடையாளப்படுத்தப்படும்.

சென்னையில் <NNP> காண <VM> வேண்டிய <VAUX> இரண்டாவது <QO>

இடம் <NN> மெரீனா <NNP> கடற்கரை <NN>ஆகும் <VAUX>. <SYM>

16. INFT (Inensifier, மிகையடை)

மிகையடை மொழியின் பெயரடைகளையுடனும் வினையடைகளுடனும் பயன்படுத்தப்படுகின்றது. எல்லா மிகையடைகளும் INFT என அடையாளப்படுத்தப்படும்.

மிகப் <INTF> பெரிய <JJ> கோபுரம் <NN>

பஸ் <NN> மிக <JJ> வேகமாகச் <RB> சென்றது <VM>. <SYM>

17. NEG (Negation, எதிர்மறை)

இல்லை போன்ற எதிர்மறைச் சொற்கள் NEG என அடையாளப்படுத்தப்படும்.

அங்கு <NST> குளிப்பதற்கு <VM> இடம்<NN> இல்லை <NEG>. <SYM>

18. SYM (Speical Symobl, சிறப்புக் குறியீடு)

பிற அடையாயாளங்களால் பாகுபடுத்தப்பட இயலாத சொற்கள் SYM என் அடையாளப்படுத்தப்படும். $ போன்ற சிறப்புக் குறியீடுகள் SYM என அடையாளப்படுத்தப்படும்.

19. RDP (Reduplication, இரட்டிப்பு)

இந்திய மொழிகளில் சில சொற்கள் இரண்டு தடவை எழுதப்படுகின்றன. அம்மாதிரியான சொற்கள் RDP என அடையாளப்படுத்தப்படும்.

யாத்திரிகள் <NN> மலையில் <NN> மெல்ல மெல்ல <RDP> அடிவைத்து <MV> ஏறினார்கள் <VM>. <SYM>

20. ECH (Echo Words, எதிரொலிச் சொல்)

எதிரொலிச் சொற்கள் இந்திய மொழிகளில் பொதுவாகக் காணப்படும்.

அவர்கள் <PRN> தடதட <ECH> என்று படிகளில் <NN> ஏறினார்கள் <VM>.

21. UNK (Unkown Words, தெரியாத சொற்கள்)

தெரியாத சொற்கள் UNK என அடையாளப்படுத்தப்படும்.

4.7.2 சொல்வகைப்பாடு அடையாளப்படுத்தும் பொறியின் செயல்பாட்டுத் தனிக் குறிப்பீடுகள் (Functional Specifications of POS tagging Engine)

வாக்கியங்களில் வரும் ஒவ்வொரு சொல்லுக்கும் சொல்வகை அடையாளப்படுத்தும் செயல்பாடு சொல்வகை அடையாளப்படுத்தல் (Parts of Speech (POS) tagging) எனப்படும். வாக்கியங்களின் வரும் உறுப்புகளின் பங்களிப்பை ஆய்வதற்குப் பெயர், வினை, பெயரடை, வினையடை போன்ற சொல்வகைப்பாடுகளைக் கண்டுபிடிப்பது உதவும். இதற்கு விதி அடிப்படையிலானது (rule-based), புள்ளியியல் அடிப்படையிலானது (statistics based), மாற்றம் அடிப்படையிலானது (transformation based) எனப் பல அணுகு முறைகள் இருக்கின்றன. இ-இ இயந்திர மொழிபெயர்ப்பில் புள்ளியியல் அணுகு முறை பயன்படுத்தப்படுகின்றது.

4.7.2.1 இடுவரல் விடுவரல் தனிக்குறிப்பீடுகள்

Input TKN_

Output CAT_

CAT_ என்ற தனிதன்மையின் எல்லாச் சாத்தியங்களான மதிப்புகளும் இங்கு பட்டியலிடப்படவேண்டும். NN, PRP, VM மற்றும் பிற.

இடுவரல் தனிக்குறிப்பீடுகள் TKN_ என்ற தனிதன்மை இடுவரல் சொல்வகைப்பாடு அடையாளப்படுத்திக்குத் தரப்படும் SSF-இல் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும்.

விடுவரல் தனிக்குறிப்பீடுகள் CAT_ என்ற தனித்தன்மை சொல்வகைப்பாடு அடையாளப்படுத்தியால் வரையறை விளக்கம் செய்யப்படவேண்டும் என வேண்டும். எனவே விடுவரல் SSF சில செல்லுபடியாகும் மதிப்புகளுக்கு CAT தனிப்பண்பைக் கொண்டிருக்கவேண்டும்.

CAT_ என்பது PRP, NN போன்ற மதிப்புகளைக் கொண்டிருக்கும்.

எல்லா CAT_ மதிப்புகளையும் இங்கு பட்டியலிடவும்; பின்னர் ஒவ்வொன்றிற்கும் ஒரு சுருக்கமான வருணனையைத் தரவும்.

PRP - Pronoun

NN – Noun

எடுத்துக்காட்டு:

Input

ADDR_	TKN_	OTHR
1	Afwa	<fs af =’ afwa pronn…’>
2	pEyanE	<fs af =’ pEyanE, noun…’>
3	Pola	<fs af =’ pola, psp…’
4	irukkirYAn	<fs af =’ irukkiRaan, verb…’

Output

ADDR_	TKN_	CAT_	OTHR
1	afwa	PRP	<fs af =’ afwa, pron…’
2	pEyanE	NN	<fs af =’ pEyanE, noun…’
3	pola	PSP	<fs af =’ pola, psp…’>
4	irukkirYAn	Verb	<fs af =’ irukkirYAn, verb…’

(குறிப்பு: நீங்கள் ஒரு பண்புக்கூறை (feature) பயன்படுத்துவதானால் இடுவரல் மாதிரியிலும் விடுவரல் மாதிரியிலும் காட்டவும்.)

4.7.2.2 சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம்

சொல்வகைப்பாடு அடையாளப்படுத்தியின் ஒழுக்குப் படம் ஒரு தொகுதியிலிருந்து மற்றொரு தொகுதிக்கு வழியமைப்புக் கட்டுப்பாட்டு ஒழுக்கின் (program control flow) முழுக் காட்சியையும் காட்டும். இது வழியமைப்புக்குள் பல்வேறுபட்ட தீர்மான இடங்களைக் காட்டும்.

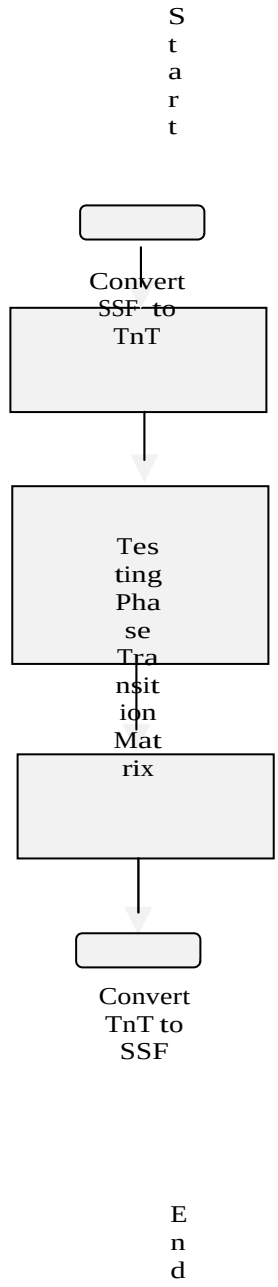


Fig 28 — Flow chart for POS Tagger

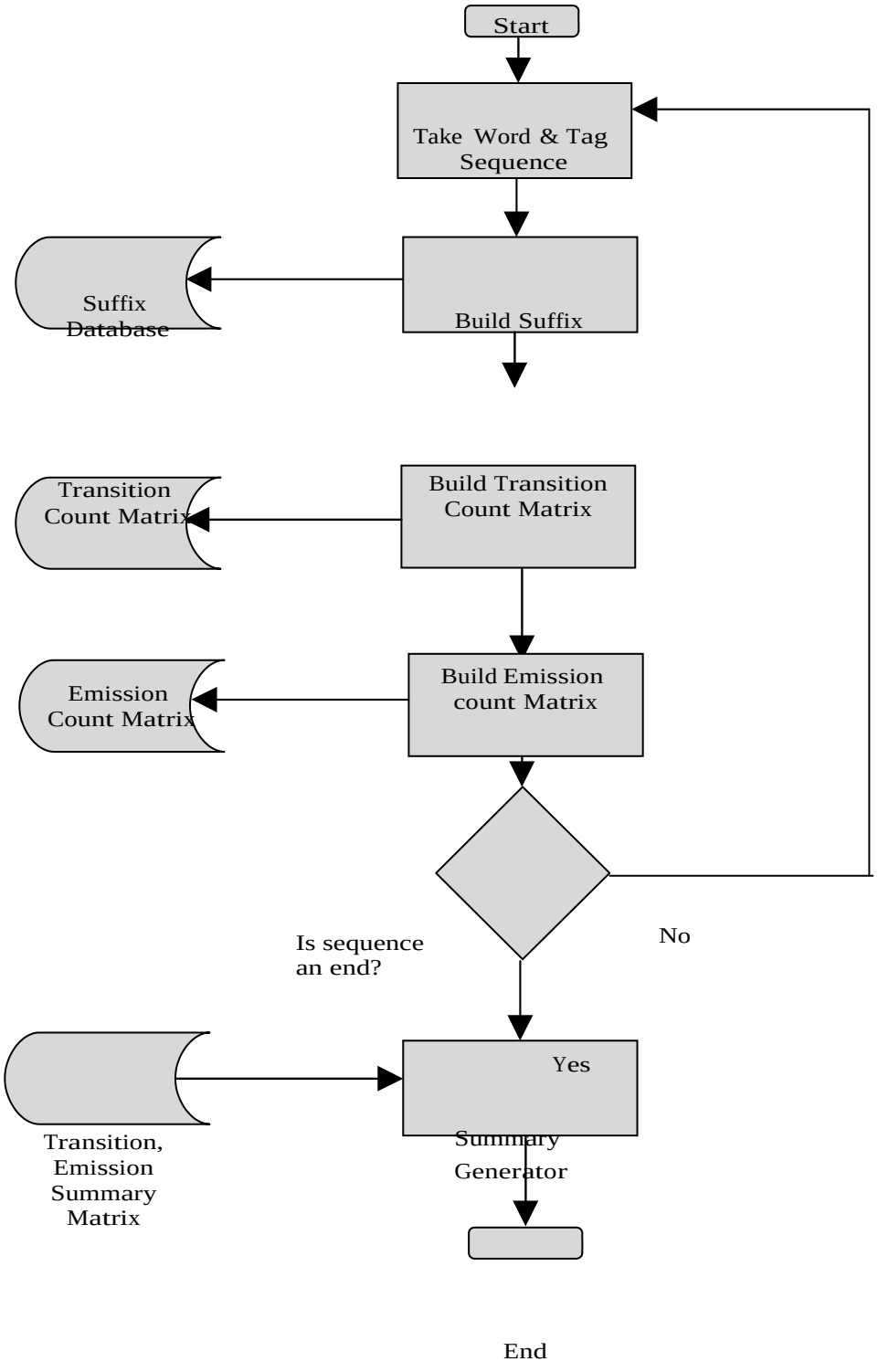


Fig.29 — Flow Chart of Training Phase

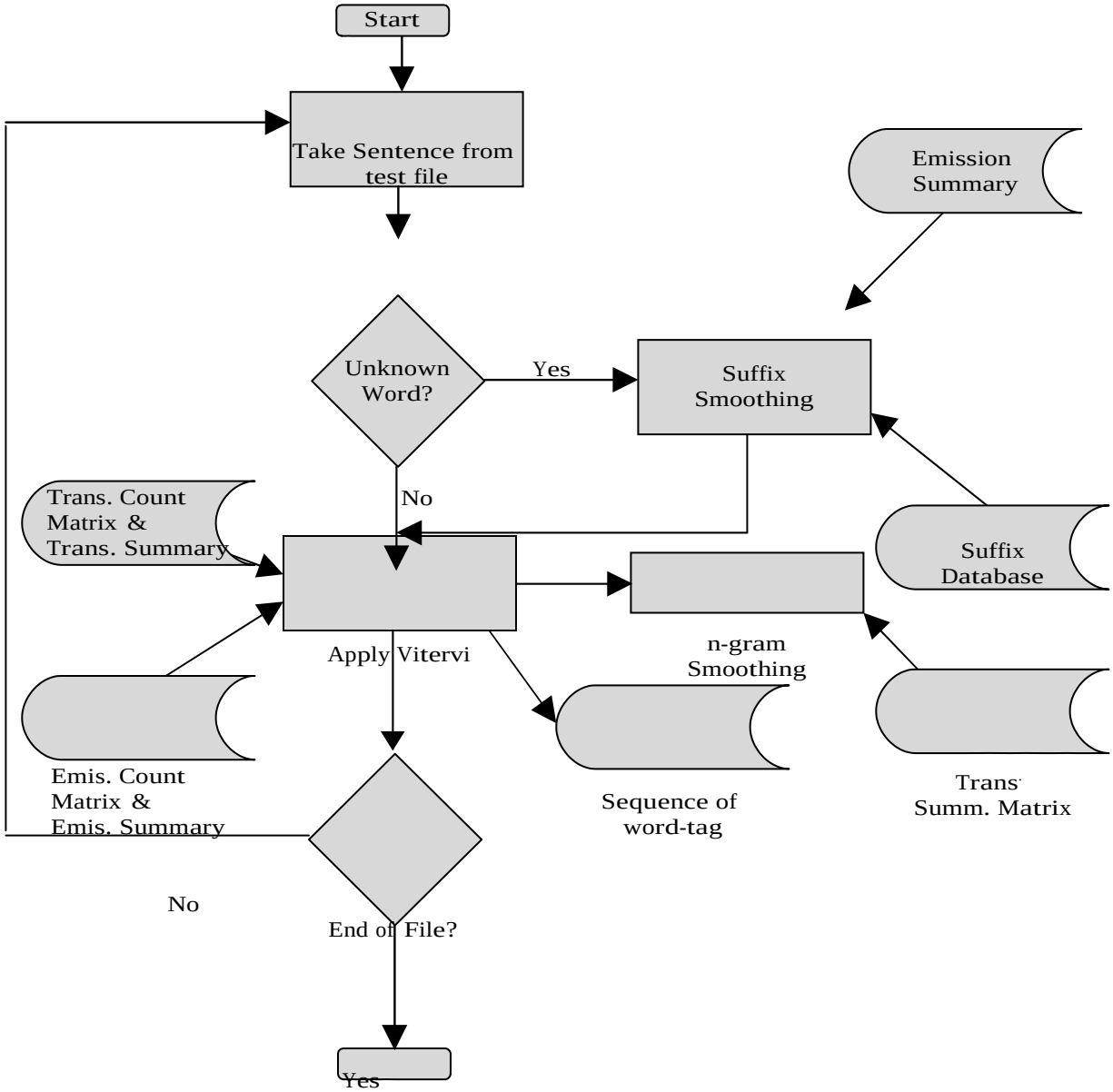


Fig.30 — Flow Chart of Testing Phase

4.7.2.3 செயல்பாட்டு வருணனைகள்

1) SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று (Covert SSF to TnT Format)

   பயிற்சிக் கோப்பை SSF வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று. மேலும் சோதனை கோப்பையும் வடிவமைப்பிலிருந்து TnT வடிவமைப்புக்கு மாற்று.

2) TnT வடிவமைப்பிலிருந்து SSF வடிவமைப்புக்கு மாற்று (Convert TnT Format to SSF)

   TnT வடிவமைப்பில் உள்ள சொல்வகைப்பாடு அடையாளப்படுத்தியால் உருவாக்கப்பட்ட இடுவரலை SSF வடிவமைப்புக்கு மாற்று.

3) எழுத்துக் கிளையை உருவாக்கு (Build Letter Tree)

   முன்னொட்டு கிளைத் தரவு அமைப்பை உருவாக்கப் பயிற்சிக் கோப்பிலிருந்து சொற்களையும் அடையாளங்களையும் எடுக்கவும். கிளையமைப்பில் சொல்லையும் அடையாள நிகழ்வெண்ணையும் சேகரிக்கவும். சொல்லையும் அதன் நிகழ்வெண்ணையும் பங்கெடுப்பாளர்களாகக் எடுத்துகொண்டு எழுத்துக் கிளையை உருவாக்கவும்.

4) புடைபெயர்வு எண்ணிக்கை சட்டத்தைக் உருவாக்கவும் மற்றும் வெளிவரு எண்ணிக்கை சட்டத்தையும் உருவாக்கவும் (Build Transtition Count Matrix and Build Emission count matrix)

   அடையாள வரிசை மற்றும் அதன் நிகழ்வெண்ணின் கலவையை உருவாக்கவும்.

5) மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடித்தல் (Finding Best Tag Sequence)

   ஒரு வாக்கியத்தை எடுக்கவும் மிகச் சிறந்த அடையாள வரிசையைக் கண்டுபிடிக்க Veterbi algorithm-ஐப் பயன்படுத்தவும்.

6) அடையாள வரிசையைக்கு மெருகிடுதல் (Smoothing for Tag sequence)

ஒரு அடையாள வரிசை இல்லாவிட்டால் சொல்வகைப்பாட்டின் பங்கெடுப்பாளர் அடிப்படையில் மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்.

7) சொற்களுக்கு மெருகிடுதல் (Soothing for Words)

சட்டகங்களில் சொல் கண்டுபிடிக்கப்படாவிட்டால் பின்னொட்டு மெருகிடும் தொழில்நுட்பத்தைப் பயன்படுத்தவும்

8) N-gramms-க்கு மெருகிடுதல் (Soothing for N-grams)

இயல் 5

தமிழ்த் தரவுத்தொகுதியில் பெயர்த்தொடர் உறுப்பு வரையறுத்தல்

5.1 முன்னுரை

பெயர்த்தொடர் உறுப்பு வரையறுத்தல் என்பது வாக்கியங்களைத் தொடர்களாகப் பகுத்துக் குறித்தலின் முக்கியக் கூறாகும். பகுத்துக்குறித்தல் என்பது ஒரு இலக்கண அடிப்படையில் பனுவல்களின் தானியக்க ஆய்வுகளுடன் தொடர்புடையது. தொழில் நுட்ப அடிப்படையில் இது ஒரு பனுவலுக்கு தொடரியல் அமைப்பைத் தரும் செற்பாங்கைக் குறிப்பிடுகின்றது. இது பெரும்பாலும் அடிப்படை உருபன்-தொடரியல் வகைப்பாடுகள் கண்டுப்பிடிக்கப்பட்ட பின் நடைமுறைப்படுத்தப்படும். வேறுபட்ட இலக்கணங்கள் (சார்பு இலக்கணம், சூழல் இலக்கணம் போன்றவை) அடிப்படையில் பகுத்துக் குறித்தல் ஒன்றுடன் ஒன்று உயர்ந்த நிலைத் தொடரியல் உறவுகளில் கொண்டு வருகிறது. வாக்கிய நிலைப் பகுத்துக் குறித்தல் சொல்நிலைப் பகுப்பாய்விலிருந்து பெறப்பட்ட தகவலைப் பயன்படுத்த தானியக்கச் சூழல் அடிப்படையிலான மற்றும் சூழல் சுதந்திரமான தொடரியல் ஆய்வை உள்ளடக்குகின்றது. ஒரு பகுத்து அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி பகுத்துக் குறித்தலில் கிளைப்படங்களைப் பயன்படுத்துவதால் இது கிளைவங்கி (Tree bank) என்றழைக்கப்படுகின்றது. கிளை அமைப்பின் காட்சிப்படம் மிக அரிதாகவே தரவுத்தொகுதி அடையாளப்படுத்தலில் காணப்படும். பொதுவாக ஒத்த தகவல் அடையாளப்படுத்தப்பட்ட அடைப்புக்குறிகளின் குழுமத்தைப் பயன்படுத்தி உருப்படுத்தம் செய்யப்படும்.

எ.கா.

'Pearl sat on a chain' என்பது கிளை வங்கியில் பின்வருமாறு தோன்றும்.

[S [NP Pearl – NP/NP] [VP – Sat VVI] [PP on II]

[NP a – AJI] [Chair – NNI]

[NP] [PP] [VP] S]

இதில் உருபனியல், தொடரியல் தகவல் கீழ்க்கோடுகளால் சொற்களுடன் இணைக்கப்பட்டிருக்கும் உறுப்புகள் தொடர்களின் தொடக்கத்திலும் இறுதியிலும் திறக்கும் மற்றும் அடைக்கும் செவ்வக அடைப்புக் குறிகளால் அடையாளப்படுத்தப்படும்.

எல்லாப் பகுத்துக் குறித்தல் ஒழுங்கு முறைகளும் ஒன்றல்ல. முக்கியமான வேறுபாடுகள் பின்வருமாறு:

1. ஒரு ஒழுங்குமுறை பயன்படுத்தும் உறுப்பு வகைகளின் எண்ணிக்கை.
2. இந்த உறுப்பு வகைகள் ஒன்றுடன் ஒன்று இணைவதற்கு அனுமதிக்கப்படும் வகைகள்

இந்த வேறுபாடுகள் இருப்பினும் பெரும்பாலான பதுத்துக் குறித்தல் திட்டங்கள் சூழல் கட்டுப்பாடு இல்லாத தொடரமைப்பு இலக்கணத்தின் அடிப்படையால் அமைந்தது.

இந்த அமைப்பொழுங்குக்குள் முழு பகுத்துக் குறித்தல் [full parsing] திட்டம் வாக்கிய அமைப்பின் விரிவான ஆய்வைத் தருவதை நோக்கமாகக் கொண்டது. குறுகிய பகுத்துக் குறித்தல் திட்டம் (Skeleton parsing scheme) தொடரியல் உறுப்பு வகைகளில் குறைந்த நிலையிலான வேறுபடுத்தலைப் பயன்படுத்துகின்றது. மேலும் சில உறுப்பு வகைகளின் அக அமைப்பை விட்டுவிடுகின்றது. பகுத்துக் குறித்தல் பெரும்பாலும் மனித ஆய்வில் பின் திருத்தம் செய்யப்படுகின்றது. ஏனென்றால் தானியங்கு பகுத்துக் குறித்தல் சொல்வகை அடையாளப்படுத்தலை விட வெற்றி குறைந்தது. முழு மனித இயக்கப் பகுத்துக் குறித்தலின் குறைப்பாடு பகுத்துக் குறித்தல், தரவுத்தொகுதி திரித்தல் இவற்றில் ஈடுபடும் ஆய்வாளரின் ஒழுங்கின்மையாகும். இதை ஈடுக்கட்ட விரிவான வழிமுறைகள் தரப்படுகின்றன. இருப்பினும் பல்பொருள் கோள்

சாத்தியமாகும்போது மயக்கம் ஏற்படலாம். கிளை வங்கிகள் இயற்கை மொழிகளுக்கு அதன் அமைப்பின் வேறுபட்ட நிலைகளில் (சொல்நிலை, வாக்கியநிலை, தொடர்நிலை, செயல்பாடு–பங்கெடுப்பாளர் நிலை) அடையாளப்படுத்தும் நிலை தரும் மொழி மூலவளமாக அமைகின்றது. கிளை வங்கிகள் இயற்கைமொழி ஆய்வு தரவு உந்தல் அடிப்படைகள், மனித மொழி தொழில் நுட்பங்கள், இலக்கணப் பிரித்தெடுத்தல், பொதுவாக மொழியியல் ஆய்வுகள் இவற்றின் உருவாக்கத்திற்கு முக்கியமானதாகும்.

அடிப்படையில் தொடர் பகுப்பான் (chunker) வாக்கியத்தை மேலுறல் செய்யாத முக்கியமான தொடர்களாகப் பிரித்து அதற்குப் புலக்குறிப்பு (label) தருகிறது. தொடர்பகுப்பான் அவற்றின் துல்லியமான வெளிப்யீடு அடிப்படையிலும் எவ்வாறு தொடர் பகுதிகள் விளக்கப்படுகிறது என்பதன் அடிப்படையிலும் வேறுபடுகின்றன. பல தொடர் பகுப்பான்கள் எளிய தொடாகளைப் பகுத்தலைவிடக் கூடுதலாகச் செய்கிறது. பிற பெயர்த்தொடர்களை மட்டும் கண்டுபிடிக்கிறது. தொடர்பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலுக்கும், முழு பகுத்துக் குறித்தலுக்கும் இடையில் வருகிறது. தனிப்பட்ட தொடர் பகுதிகளின் அமைப்பை விளக்குவது எளிது. ஆனால் அவற்றிற்கிடையிலான உறவுகளை விளக்குவது கடினம். இது தனிப்பட்ட சொல் வகுப்புகளைச் சார்ந்திருக்கிறது. இவ்வாறு தொடர்களைப் பகுத்தில் தற்பொழுது இருக்கின்ற முறைக்கும் முன்மாதிரி வெளிப்யீட்டு ஆய்வுக்கும் இடையில் உள்ள சலுகையான முறையாகும். தொடர் பகுப்பான்கள் வாக்கியங்களைப் பிரித்து அடையாளப்படுத்துகிறது. பெரும்பாலான தொடர் பகுப்பான்கள் சொல்வகை அடையாளப்படுத்தலில் உள்ள தகவலைப் பயன்படுத்துகிறது. பிற உண்மையான சொற்களைப் பார்க்கிறது.

சொல் வகைப்பாட்டு அடையாளப்படுத்தி சொல் மற்றும் சூழலுக்கான தகவலைப் பயன்படுத்தி இம்மயக்கங்களை நீக்குகின்றது. சொல் வகைப்பாட்டுத்

தகவல் பகுத்துக்குறித்த அமைப்பில் உறுப்பின் நிரல் அல்லது இடத்தைக் கண்டுகொள்ள தேவைப்படுகிறது. பண்புக்கூறு தகவல்கள் நிரப்பியையும் உடன்பாட்டையும் பரிசோதிக்கப் பயன்படுகிறது.

தமிழில் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமானது. தொடரமைப்பு விதிகளின் எல்லா விதமான வரிசை மாற்றங்களுக்கும் வேண்டி எழுதப்பட வேண்டும் என்று அவசியப்படும் இவ்வகங்களின் காரணமாக இப்பகுத்துக் குறிப்பான்கள் தேவையில்லாமல் விதிகளின் எண்ணிக்கையைப் பெருக்கிக் கொள்ளும். பகுத்துக்குறிப்பான் மீது கூடதலான சுமை புகுத்தப்படும். தற்போதைய ஆய்வில் பகுத்துக்குறித்தலுக்கு முன் வாக்கியத்தை ஒரு குறிப்பிட்ட நிலை பேறாக்கம் செய்யும் நடத்தை பின் பற்றப்படுகிறது.

வினைதான் வாக்கியத்தின் முக்கிய உறுப்பாகும். வினைகள் அவற்றின் பொருண்மை இயல்பு அடிப்படையில் வேறுபட்ட பங்களிப்பாளர் அமைப்பை ஏற்கும். இப்பங்கெடுப்பாளர் அமைப்புகள் பெயர்த்தொடர்களின் வேற்றுமை உருபுகளை விளக்கும். அவை வினைக்கு நேரடியான மற்றும் நேரடியல்லாத செயப்படுபொருளாக அமையும். பகுத்துக்குறிப்பான் வாக்கியத்தில் வினைக்குப் பொருத்தமான பங்கெடுப்பாளர் அமைப்பை பரிசோதிப்பதைக் கவனிக்க வேண்டும். மேலும், தமிழில் எழுவாயின் பால்-எண்-இட பண்புக்கூறுகள் பயனிலையுடன் இயைபுடையதாக இருக்க வேண்டும். ஒரு பகுத்துக்குறிப்பான் வாக்கியத்தைப் பகுத்துக்குறிக்கும்போது வாக்கியத்தைச் சரியாக மதிப்பிட இவ் இயைபு பண்புக்கூறைப் பரிசோதிக்க வேண்டும்.

5.2 பகுத்துக்குறித்தல்

பகுத்துக்குறித்தல் இயந்திரமொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுத்தல் போன்ற பயன்பாடுகளில் முக்கியமான செயல்பாடாகும். சில பயன்பாடுகள் வாக்கியங்களின் முழுமையற்ற பகுத்துக்குறித்தலை (shallow parsing) வேண்டுகிறது. ஆங்கிலம் மற்றும் பிற ஐரொப்பிய மொழி வாக்கியங்களின்

பகுத்துக்குறித்தலுக்குப் பல முயற்சிகள் மேற்கொள்ளப்பட்டுள்ளன. இந்திய மொழிகளில் இந்தி மற்றும் வங்காள மொழிகளுக்கு பகுத்துக்குறிப்பான்கள் இருந்தாலும் தமிழைப் பொறுத்தவரையில் கூறத்தக்க பகுத்துக்குறிப்பான் இல்லை. தமிழில் பல கோட்பாட்டு மொழியியல் அடிப்படையிலான தொடரியல் ஆய்வுகள் செய்யப்பட்டுள்ளன. இருப்பினும், தமிழ் வாக்கியங்களின் பகுத்துக்குறித்தலுக்கு வேண்டிய கணினியியல் சார் சட்டகத்தை உருவாக்க அச்சிட்டு வெளிப்யிடப்பட்ட புத்தகங்கள் இல்லை. தற்போதைய ஆய்வு தமிழ் வாக்கியங்களின் பகுத்துக் குறித்தலுக்கு கணினி அடிப்படையிலான இலக்கணச் சட்டகம் உருவாக்குவதில் கவனம் செலுத்தப்படுகின்றது.

5.3 ஆழமிலாப் பகுத்துக் குறித்தல்

நாம் தொடரியல் என்ற கலைச்சொல்லை மரபுப் பகுப்பாய்விலிருந்து கிடைக்கும் வெளிப்யீட்டை விடக் குறைவான நிறைவுடைய ஆய்வைக் குறிக்கும் பொதுவான சொல்லாகப் பயன்படுத்துகிறோம். ஆய்விலிருந்து கிடைக்கும் வெளிப்யீடு தொடரமைப்புக் கிளைப்படம் அல்ல. ஒரு பகுப்பாய்வி பெயர்த்தொடர்கள் போன்ற பகுக்கப்பட்ட உறுப்புகளை வாக்கியத்தில் அவற்றின் அக அமைப்பையும், செயல்பாட்டையும் குறிப்பிடாமல் அடையாளம் காணும். மற்றொரு வகையான ஆய்வு முதன்மை வினை மற்றும் அவற்றின் நேரடியான பங்கெடுப்பாளர் போன்ற சொற்களின் சில செயல்பாட்டுப் பங்களிப்பை அடையாளம் காணும். பகுப்பாய்வு ஒழுங்குமுறை உருபயனியல் ஆய்வின் மீதும் பொருண்மை மயக்க நீக்கலின் மீதும் இயல்பாக வேலை செய்யும். இதன் அடிப்படை நோக்கம் தலைச்சொல் (Head word) உருபனியல் தகவல் மற்றும் சொல்வரிசை அமைப்பு இவற்றில் இருந்து சாத்தியமான தொடரியல் அமைப்பைப் (Syntantic structure) பெறுவதாகும். முன்மாதிரியாக பகுப்பாய்வுத் தொடர்களையும் தலை-அடை (head-medifier) உறவுகளையும் கண்டுபிடிப்பதை நோக்கமாகக் கொள்கிறது. பெரிய உரை தரவுத்தொகுதிகளின் (corpus) பயன்பாட்டில் ஆழமில்லாப் பகுப்பாய்வி பயன்படுத்தப்படுகிறது. பெரும்பாலும் பகுப்பாய்விற்கு எல்லாச் சிக்கல்களுக்கும் தீர்வு காண்பதற்கு போதிய திறன்

இல்லாவிட்டால் பகுதியான ஆய்வுகள் அனுமதிக்கப்படும். சர்ச் (Church.1988) பொருத்தமான அடைப்புக் குறிகளைச் செருகி எளிய பெயர்த் தொடர்களை அடையாளம் காணும் புள்ளியல் (Stochastic) வழியமைப்பை வடிவமைத்துள்ளார். ஒப்னெய் (Obney, 1991) உளமொழியியல் சான்றுகள் அடிப்படையிலும், நடைமுறை பயன்பாடுகள் அடிப்படையிலும் ஆழமில்லாப் பகுப்பாய்வின் முக்கியத்துவத்தை முதலில் விவாதித்தவர் என்று மதிக்கத்தக்கவர் ஆவார். அவர் தமது அணுகுமுறையில் ஆழமில்லாப் பகுப்பாய்வைப் பெறுவதற்காக முற்றுநிலை மாறிகளைப் பயன்படுத்தினார். ஆழமில்லாப் பகுப்பாய்வி சொல்வகைப்பாடு அடையாளப்படுத்தல், தொடர் உறுப்பு வரையறுத்தல், தொடர் உறவுகளைக் கண்டுபிடித்தல் என்ற முன்மாதிரி தொகுப்புகளை உள்ளடக்கியது.

5.3.1 சொல்வகை அடையாளப்படுத்தல்

ஒரு சொல்லும் அதன் சூழலும் தரப்படுகையில் சொல்வகை அடையாளப்படுத்தி அச்சொல்லின் சரியான உருபனியல்-தொடரியல் வகுப்பை (பெயர், வினை கொண்ட பகுப்பை) உறுதிசெய்யும் சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தல் (Parts of Speech tagging) இயற்கைமொழி ஆய்வில் மிக அறியப்பட்ட சிக்கலாகும். இதில் இயந்திர கற்றல் அணுகுமுறைகள் (machine language approaches) பொதுவாகப் பயன்படுத்தப்படுகிறது.

5.3.2 தொடர் உறுப்பு வரையறுதல்

சொற்களும் அவற்றின் உருபனியல்-தொடரியல் வகுப்புகளும் தரப்படுகையில் எந்தச் சொற்கள் தொடர் உறுப்புகளாகக் (பெயர்த்தொடர், வினைத்தொடர், நிரப்பியத்தொடர் போன்றவைகளாக) குழுமப்பட வேண்டும் என்று தொடர் உறுப்பு வரையறுத்தல் (Chunking) செயல்பாட்டால் தீர்மானிக்கப்படுகிறது.

தொடர் உறுப்பின் வரையறை விளக்கத்துடன் பின்வரும் கேள்வி தொடர்புள்ளது. தொடர் உறுப்பு எதைக் கொண்டிருக்கும்? ஒரு மாதிரி தொடர் உறுப்பு ஒரு தனி நிலைப் பொருளடக்கச் சொல்லைச் சுற்றி ஒரு கூட்டம் செயல் பாட்டுச் சொற்களைக் கொண்டிருக்கும் (Abney, 1991). தொடர் உறுப்புகள் இயல்பாக ஒன்றுக்கொன்று இயைபுள்ள சொற்களின் குழுமமாக எடுத்துக்கொள்ளப்படுகின்றது.

அடுத்த கேள்வி 'நம்முடைய தேவைக்காக இந்த இயைபுள்ள சொற்களின் குழுவின் எல்லைகளை எவ்வாறு வரையறை விளக்கம் செய்யவியலும்?' என்பதாகும். எடுத்துக்காட்டாக, பின்வரும் இணைகளில் எது சரியான தொடர் உறுப்பு? என்ற ஐயம்.

1. ((மேசையின் மேல்)) அல்லது
2. ((மேசையின்)) ((மேல்))

5.3.3 தொடர் உறவைக் கண்டுபிடித்தல்

ஒரு வாக்கியத்தில் உள்ள தொடர் உறுப்புகள் தரப்படுகையில் முக்கிய வினையுடன் அவற்றிற்குள்ள உறவுகள் (எழுவாய், பயனிலை, இடம் போன்றவை) தீர்மானிக்கப்பட வேண்டும் என்ற நிலையில் உறவைக் கண்டுபிடித்தல் (Relation finding) முக்கியமான செயல்பாடாகும்.

ஆழமில்லாப் பகுப்பாய்விகள் இயற்கை மொழி முழுவதையும் கையாள வேண்டி இருப்பதால் அவை பெரிதாகவும் பெரும்பாலும் ஆயிரக்கணக்கான விதிகளையும் கொண்டிருக்கின்றன. இவ்விதிகளில் பல விதிவிலக்கானவை என்பதன் அடிப்படையில் இவை வலிமை குறைந்தவையாக இருக்கின்றன.

ஆழமில்லாப் பகுப்பாய்விகள் இயந்திர கற்றல் அடிப்படையிலான தொழில்நுட்பங்களைப் (techniques) பயன்படுத்தித் தானியக்கமாக உருவாக்கப் படுகின்றன.

5.4 பெயர்த்தொடர் உறுப்பு வரையறுத்தல்

பெயர்த்தொடர் உறுப்பு வரையறுத்தல் (பகுத்தல்) வாக்கியங்களிலிருந்து பெயர்த் தொடர்களைப் பிரித்தெடுப்பதைக் குறிப்பிடுகின்றது. பகுத்துக்குறித்தலைவிட பெயர்த்தொடர் பகுத்தல் எளிதாக இருந்தாலும் ஒரு துல்லியமான மற்றும் திறமைவாய்ந்த பெயர்த்தொடர் பகுத்தல் பல பயன்பாடுகளில் பயன்படுத்தப்படுகின்றது என்ற உண்மையால் மிக முக்கியத்துவம் வாய்ந்ததாகக் கருதப்படுகின்றது.

பெயர்த்தொடர்களைப் பனுவல்களைப் பகுத்துக் குறிப்பதற்கு முன் முன்பகுப்பாய்வுக் கருவியாகப் (pre-processing tool) பயன்படுத்த இயலும். இயற்கை மொழியின் கூடுதலான பொருண்மை மயக்கம் காரணமாக பனுவலின் சரியான பகுத்துக்குறித்தல் மிகக் கலவைத் தன்மையானதாக (சிக்கலானதாக) இருக்கின்றது. இந்நேர்வுகளில் தொடர்களைப் பகுத்தலை இம்மயக்கங்களை ஓரளவு தீர்ப்பதற்கான முன்பகுப்பாய்வுக கருவியாகப் பயன்படுத்த இயலும். இப்பயன்பாட்டில் தொடர் பகுத்தலைச் சொற்களுக்குப் பதிலாகத் தொடர் பகுதிகளைச் சார்ந்திருக்கும் ஆவணங்களிலிருந்து தரவை மீட்பதற்குப் பயன்படுத்த இயலும். குறிப்பாகப் பெயர்களும் பெயர்த்தொடர்களும் தகவல் மீட்புக்கும் பிரித்தெடுத்தலுக்கும் மிகவும் பயனுள்ளதாகும். இயந்திர மொழிபெயர்ப்பு மீதான சமீபகாலச் செயல்பாடுகள் பயனுள்ள மாற்ற அமைப்பொழுங்குகளை (transfer patterns) ஆக்குவதற்கு இரு மொழிகளிலுள்ள பனுவல்களை (இணை தரவுத்தொகுதிகளை) (parallel corpora) பயன்படுத்துகின்றது. பெயர்த்தொடர்கள் இணை தரவுத்தொகுதிகளில் உள்ள பனுவல்களைப் பொருத்தும் பயன்பாடுகளில் பயன்படுகின்றது. இணை தரவுத்தொகுதிகளில் உள்ள வாக்கியங்களைத் தொடர்ப்பகுப்புத் தகவல்களைப் பயன்படுத்தியும் மூல மொழியிலுள்ள தொடர்

பகுதிகளை இலக்கு மொழியிலுள்ள தொடர் பகுதிகளுடன் உறவுபடுத்தியும் பொருத்த வரைவு செய்யலாம். மேலும் பகுக்கப்பட்ட பெயர்த் தொடர்களைத் தரவின் தரமான பகுத்துக்குறித்தல் தேவைப்படாத பிற பயன்பாடுகளிலும் பயன்படுத்த இயலும்.

5.5. தமிழ்த் தரவுத் தொகுதியைத் தொடர் உறுப்புக்கு வரையறுத்தல்

5.5.1 தமிழ்த் தொடரியல் அமைப்பு

தமிழைப் பொறுத்தவரையில் அது ஒரு ஒட்டுக்கள் நிறைந்த மொழியாகும். இவ்வொட்டுக்கள் வேர்ச்சொற்களில் சேர்க்கப்பட்டு வேறுபட்ட சொல் வடிவுகளை உருவாக்கும். ஏனவே தமழுக்கு ஒரு உருபனியல் ஆய்வியை உருவாக்குவது ஒரு சிக்கலான செயல்பாடாகும். பல வாக்கியங்களுக்கு உருபனியல் பகுப்பாய்வியைப் பயன்படுத்தி சொற்களின் அல்லது ஒரு சொல்லின் சொல் வகைப்பாட்டையும் (Parts of speech (POS)) பண்புக்கூறு தகவலையும் அடையாளம் காண இயலும். இருப்பினும், சில சொற்களுக்கு மயக்கமுள்ள சொல்வகைப்பாடும் இலக்கண பண்புக்கூறுகளும் இருக்கும்.

தமிழ்த் தொடரியல் பகுத்துக் குறிப்பானுக்கு அடிப்படையாக அமையும் தமிழ்த் தொடரியல் குறித்து இங்கு விளக்கப்படும். தமிழ் வாக்கியங்களின் வகைகள் பற்றியும், அவற்றின் அமைப்பு குறித்தும் விளக்கப்படும்.

தமிழ்மொழி வேறுபட்ட சொல் வரிசை அமைப்பைக் கொண்ட தொடர்களைக் கொண்டு வாக்கியங்களை அமைக்கத்தக்க வகையில் அமைந்த மொழியாகும். சொற்களுக்கு இடையிலான தொடர்பினைப் பயன்படுத்தி வாக்கியங்களை இலக்கணரீதியாக அமைக்கிறோம். இலக்கணங்கள் வேறுபட்ட வார்த்தைகளுக்கு இடையே தொடர்பினை உருவாக்கிப் பொருள்தரும் வாக்கியங்களை உருவாக்க உதவுகிறது. இவ்வாறு சொற்களைக் கொண்டு

தொடர்களையும் தொடர்களைக் கொண்டு வாக்கியங்களையும் உருவாக்கும் விதிகளைக் கூறுவது தொடரியலாகும்.

ஒன்றுக்கு மேற்பட்ட சொற்களஅத்தொடர்கள் இணைந்து உருவாகும் வாக்கியங்கள், அவ்வாக்கியங்கள் இணைந்து உருவாகும் கலப்பு, கூட்டு வாக்கியங்கள் ஆகியவற்றின் கட்டமைப்பையும் அவை தரும் பொருண்மையையும் ஆராய்வது தொடரியல் எனப்படும். பொதுவாகத் தொடர்களை ஆராய்வது தொடரியல். அது தொடராக இருக்கலாம் அல்லது வாக்கியங்களாக இருக்கலாம். வாக்கியம் என்பதும் ஒருவகைத் தொடரே.

5.5.1.1 தொடர்கள்

தமிழ் வாக்கியங்களைப் தலைமையாய் வரும் சொல்வகைப்பாடுகளின் அடிப்படையில் பெயர்தொடர், வினைதொடர், பெயரடைத் தொடர், வினையடைத் தொடர், பின்னுருபுத்தொடர் என வகைப்படுத்தலாம்.

5.5.1.1.1 பெயர்த்தொடர்

பெயர்த்தொடர் என்பது ஒரு பெயர்ச்சொல்லைத் தலைமைப் பெயராகக் கொண்ட தொடர் ஆகும். தலைமைச்சொல் பெயராகவோ ஆக்கப்பெயராகவோ (எகா.மயக்கம்) தொகையாகவோ (எகா. பணப்பெட்டி) வினைப்பெயராகவோ, மாற்றுப்பெயராகவோ (எகா.அவன், நான்,நீ போன்று) தொழில் பெயராகவோ (வந்தது, வருவது, வருதல், வந்தமை, வந்தவன் போன்று) இருக்கலாம். பெயர்த்தொடரில் பெயருக்கு முன் அடைகொளி அடை, எண்ணடை, பெயரடை, அளவையடை போன்ற அடைகளும் பெயரெச்சத் தொடரும் வரலாம்.

எ.கா.

பையன், அந்தப் பையன், பெரிய பையன், இரண்டாவது பையன், இரண்டு பையன்கள், என்னுடைய பையன், நேற்று வந்த பையன்.

5.5.1.1.2 வினைத்தொடர்

ஒரு வினைச்சொல்லைத் தலைமைச் சொல்லாகக் கொண்டு உருவாகும் தொடர் வினைத்தொடர் எனப்படும். வினைத்தொடரில் வினைக்கு முன்னர் வேற்றுமைத்தொடர், பின்னுருபுத்தொடர், வினையடை, வினையெச்சத்தொடர் என்பன வரலாம்.

எ.கா.

விழுந்தேன், செய்தேன், அங்கு வந்தேன்
நன்றாக விளையாடினேன், வந்து பார்த்தேன்

போன்றவை வினைத்தொடர்களாகும்.

5.5.1.2 வாக்கிய வகைகள்

வாக்கியங்களைப் பல்வேறு அடிப்படையில் பல வகைகளாகப் பிரிப்பர் இலக்கணயலார். அமைப்பினை அடிப்படையாக வைத்துத் தமிழ் வாக்கியங்களை வினைப்பயனிலை வாக்கியங்கள் என்றும், பெயர்ப் பயனிலை வாக்கியங்கள் என்றும் பிரிப்பர். இதுபோன்று வாக்கியங்களின் சேர்க்கையை அடிப்படையாகக் வைத்து தனி வாக்கியம் அல்லது எளிய வாக்கியம் (Simple sentence) என்றும், கலவை வாக்கியம் (Complex sentence) என்றும் என்றும் பிரிப்பர். வெவ்வேறு வினைகளின் பண்பையும் அதனால் அடையும் மாற்றங்களையும் வைத்து செய்வினை வாக்கியம் (active voice) என்றும், செயப்பாட்டு வினை வாக்கியம்

(passive sentence)என்றும் பிரிப்பது உண்டு. இவை மட்டுமின்றி வேறொரு பண்பினை ஒட்டி இலக்கணங்கள் வாக்கியங்களை வினா வாக்கியம் (interrogative sentence) என்றும், செய்தி வாக்கியம் (affirmative sentence) என்றும் பிரிப்பர். இவை மட்டுமல்லாமல் இலக்கணவியலார் ஏவல் வாக்கியம், செய்தி வாக்கியம் என்றும் பிரிப்பர். தமிழில் காணப்படும் வாக்கியங்களை 1. வினைப் பயனிலை வாக்கியம், (கண்ணன் வந்தான்), 2. பெயர்ப் பயனிலை வாக்கியம் (இராமன் ஒரு ஆசிரியர்), 3. பெயரடை பயனிலை வாக்கியம் (இராமன் நல்லவன்), 4. உடைமைப் பயனிலை வாக்கியம் (இராமன் என்னுடையவன், புத்தகம் என்னுடையது) என நான்காகப் பிரித்து இனம் காண்பர் சிலர் (கோதண்டராமன்,1997) .

தொடரியலில் சில முக்கியமான அமைப்புகளை மட்டுமே குறித்துச் சொல்லுகின்ற நிலையில் இதன் கண் ஒருசில அமைப்புகள் மட்டுமே எடுத்துக் கூறப்படுகின்றன.

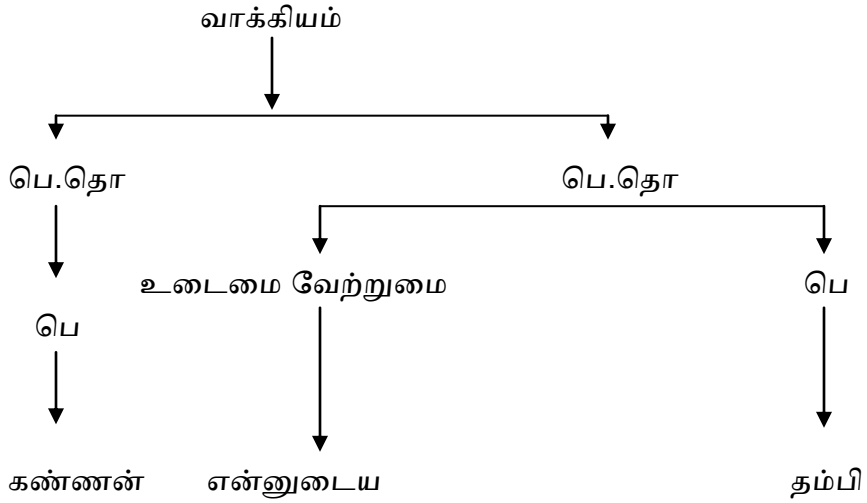
5.5.1.2..1 அடிப்படை வாக்கிய வகைகள்

தமிழில் காணப்படும் அடிப்படையான வாக்கியங்கள் 1. பெயர்ப் பயனிலை வாக்கியம் என்றும், 2. வினைப் பயனிலை வாக்கியம் என்றும் இரு வகைகளாகப் பிரிக்கலாம். இவ்விரண்டு வாக்கிய வகைகளிலும் பெயர்த்தொடர் எழுவாய்களாக வரும் இந்நிலையில் இவற்றை

1.பெயர்த்தொடர் + பெயர்த்தொடர் (NP + NP)
2.பெயர்த்தொடர் + வினைத்தொடர் (NP + VP)

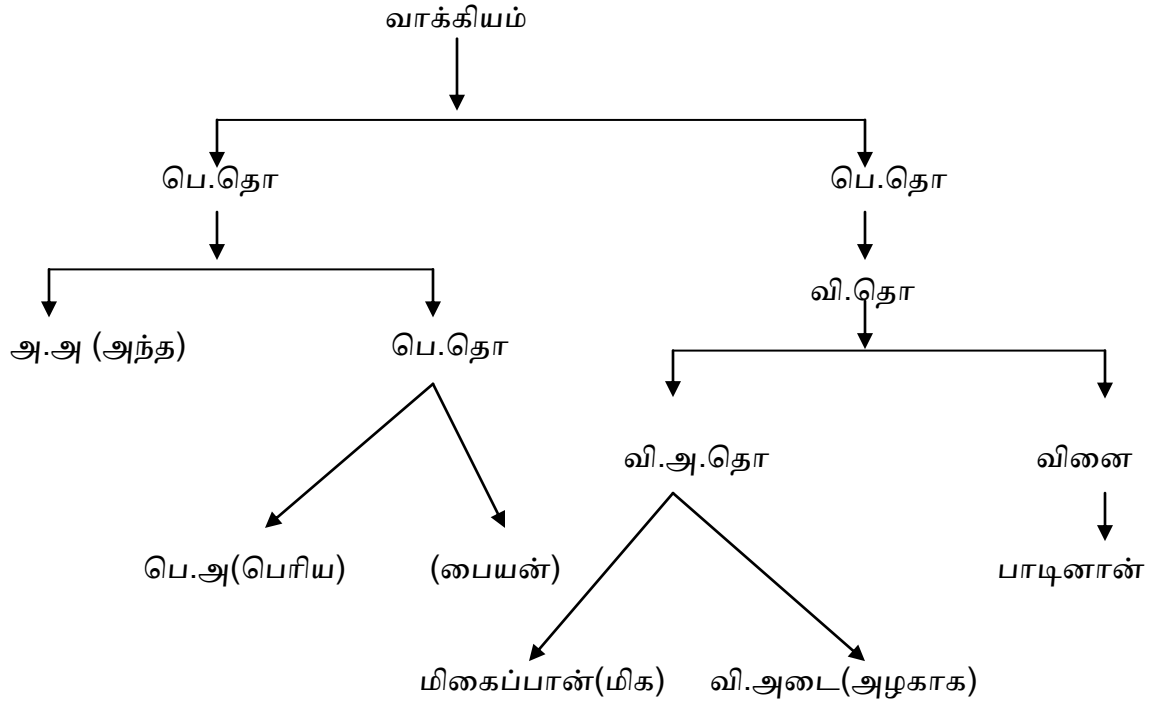
என்ற அமைப்பைக் கொண்டுள்ளவைகளாகப் பிரிக்கலாம்.

ஒரு பெயர்ப் பயனிலைத் தொடர் வாக்கியம் எவ்வாறு உள்ளது என்பது கீழே உள்ள கிளைப்படத்தில் காட்டப்பட்டுள்ளது.



மேலே ‘கண்ணன் என்னுடைய தம்பி’என்ற பெயர்ப் பயனிலை வாக்கியத்தின் அமைப்பு காட்டப்பட்டுள்ளது.

வினைப் பயனிலை வாக்கியம் என்பது ஒரு வினையையும் வினையடைத் தொடரையும், மிக என்ற மிகைப்பானையும் கொண்டு வந்துள்ளது.



5.5.1.2..2 வினா வாக்கியங்கள்

வினா வாக்கியம் என்பது தெரியாத ஒன்றை மற்றொருவரிடம் கேட்டுத் தெரிந்து கொள்ளுதற்காகப் பயன்படுத்தப்படும் வாக்கியம் ஆகும். வினாவில் அடிவினா என்றும், ஒட்டுவினா என்றும் இருவகை உள்ளன. அடிவினா யார், என்ன போன்ற வினாச் சொற்களையும் ஒட்டுவினா ஆ என்ற ஒட்டையும் (எ.கா. அவனா? இவனா?) கொண்டிருக்கும்.

வினா வாக்கியங்களும் அவற்றின் விடையாக வரும் செய்தி வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டு விளங்கும்.

ஆம் இராமன் வந்தான்.
அவன் நேற்று வந்தான்.

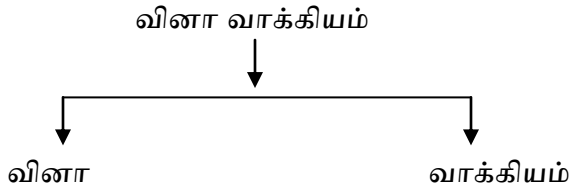
என்ற செய்தி வாக்கியங்களும்,

இராமன் வந்தான். இராமன் வந்தானா?
அவன் நேற்றா வந்தான்?

என்ற ஒட்டு வாக்கியங்களும் ஏகதேசம் ஒரே அமைப்பைக் கொண்டவை. ஒன்றில் வினா ஒட்டு உள்ளது. செய்தி வாக்கியத்தில் அது இல்லை. இதே நிலைதான் அடிவினா வரும் வாக்கியங்களிலும் காணப்படுகின்றது.

இராமன் வந்தான். இராமன் என்று வந்தான்?
இராமன் ஒரு ஆசிரியர். இராமன் யார்?
அவன் நன்கு செய்தான் அவன் எப்படிச் செய்தான்?

போன்ற இணை வாக்கியங்களும் ஒரே மாதிரியான அமைப்பையே கொண்டுள்ளன. ஒன்றில் செய்தி (நேற்று) மற்றொன்றில் வினா (என்று) இதனால் மொழியியலாளர்மள் வினா வாக்கியங்களையும் அதை ஒட்டி வரும் செய்தி வாக்கியங்களையும் ஒரே அமைப்பின் கீழ்க் கொண்டு வந்து வினா வாக்கியத்தைப் பிரித்துக்காட்டும் நிலையில் 'வினா' என்ற முன்னிலையை இணைத்து வேறுபடுத்துவர். இதனைக் கீழ வருமாறு காட்டலாம்.



ஒட்டு வினாவும், அடிவினாவும் ஒன்று வந்த இடத்தில் மற்றொன்று வருவதில்லை. இதனைக் கீழ்வருமாறு காட்டலாம்.

வினா-ஒட்டுவினா,அடிவினா
ஒட்டுவினாப் பொருள்

ஒட்டுவினா ஆம் அல்லது இல்லை என்ற பதில்களை எதிர்பார்ப்பது. முதலில் உள்ளது கேட்போரின் உள்ளத்தில் உள்ளதை உறுதி செய்வது . இரண்டாவது உள்ளதை மறுப்பது.

இராமன் வந்தான்.
இராமன் வந்தானா?

எனத் தமிழில் வருவது காணத்தக்கது. முதல் வினாவின் விடையாக ஆம் (இராமன் வந்தான்) என்றோ இல்லை (இராமன் வரவில்லை) என உறுதி செய்தோ அல்லது

மறுத்தோ விடை அமையலாம். இதுபோன்றே அடுத்த வினாவிற்கும் பதில்கள் அமையலாம்.

தமிழில் ஒட்டுவினா, வினா-ஆ என்ற உருபனால் காட்டப்படுகின்றது. பொதுவாக இது வாக்கியங்களில் ஒரே ஒரு இடத்தில் வரும் பல்வேறு இலக்கணக் கூறுகளில் ஏதாவது ஒன்றின் பின்னால் வந்து ஒட்டு வினாப்பொருளைக் காட்டும். இவ்வாறு ஏதாவது ஒன்றின் பின்னால் வருவதை வினாவின் வீச்சு (ளஉடிப்ந) என்பர் மொழியியலாளர்கள்.

இராமன் + வினா + வந்தான் >இராமன் வந்தான்.
இராமன் வந்தான் + ஆ >இராமன் வந்தானா?

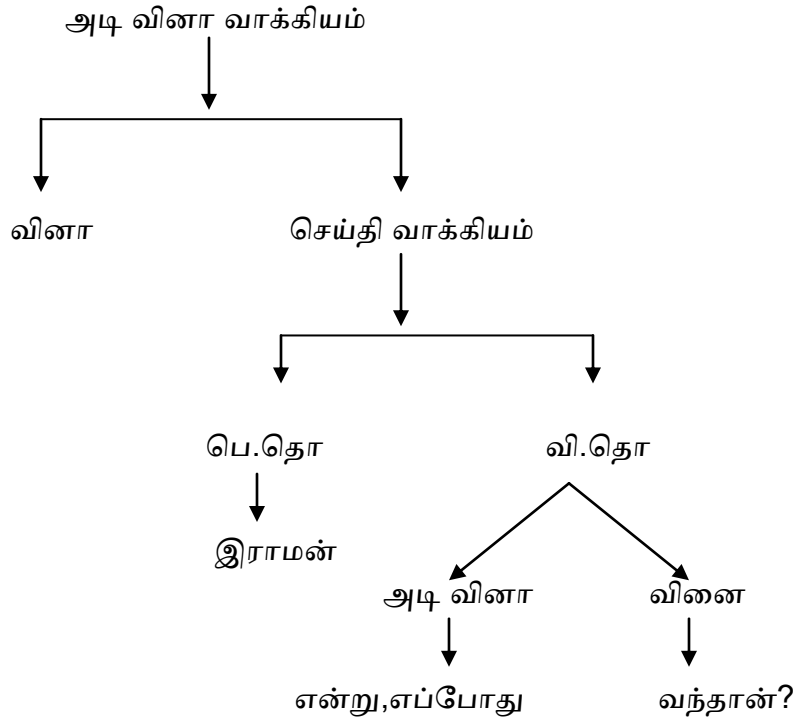
என்று வருவது குறிப்பிடத்தக்கது.

அடி வினா

அடி வினா வாக்கியங்கள் 1.யார், எவர், எவன், என்ன, எது, எவை போன்ற வினாப் பெயர்களையும், 2.எந்த, எத்தனையாவது, எத்தனை, எவ்வளவு, எப்பேர்பட்ட அல்லது எப்படி உள்ள (எத்தகைய, எம்மாதிரி போன்ற) என்ன போன்ற வினாப் பெயரடைகளையும், 3.எத்தனையாவது, எங்கு, எப்படி, எப்போது போன்ற வினா வினையடைகளையும் உள்ளடக்கியவை. இவற்றுள் சில ஒன்றுக்கு மேற்பட்ட நிலைகளிலும் (என்ன, எப்பொழுது) வரும்.

இராமன் நேற்று வந்தான்.
இராமன் என்று வந்தான்?

போன்ற வாக்கியங்கள் ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் 'நேற்று' 'என்று' போன்றவை ஒரே அமைப்பு கொண்ட வாக்கியங்களில் ஒரே இடத்தில் வருபவை ஆகும். எனவே அடி வினா வாக்கியங்களின் அமைப்பைக் கீழ்வருமாறு காட்டலாம்.



என்ற ஒரே அமைப்பைக் கொண்டுள்ள நிலையில் அடி வினா என்ற புனை நிலையில் (hypothetical) கொள்ளப்பட்டது கெட அடி வினா வாக்கியம் உருவாக்கப்பெறுகிறது.

5.5.1.2.3 செய்வினை, செயப்பாட்டு வினை வாக்கியங்கள்

வினாப் பாகுப்பாடு பற்றிக் கூறும்போது வினைகள் செய்வினை என்றும் செயப்பாட்டு வினை என்றும் பகுக்கப்படும் என்று கூறப்பட்டது. செய்வினைப்யன் செயவென் வாய்பாட்டு வினஇயெச்சத்தின் பின்னர் படு அல்லது பெறு என்ற துணை வினைகளை இணைக்கும்போது உண்டாகும் வினைதான் செயப்பாட்டு வினை எனவும் கூறப்படுகிறது. அவ்வாறு உள்ள செயப்பாட்டு வினையைப்

பயனிலையாகக் கொண்ட வாக்கியங்கள் செயப்பாட்டுவினை வாக்கியங்கள் ஆகும்.

செயப்படுபொருளை எழுவாயாக ஆக்கியும் செய்பவனை அல்லது மருத்தாவை ஆல் வேற்றுமையாக ஆக்கியும் கூறப்படுவது செயப்பாட்டு வினை வாக்கியம் (Passive Sentence).

இராஜா இராமனை அடித்தான்

என்பது செய்வினை வாக்கியம். இங்கு செய்பவன் எழுவாயாகவும், செய்யப்படுபவன் ஐ வேற்றுமையாகவும் உள்ளது. இதனை உரிய இலக்கண மாற்றத்தால் மாற்றி

இராமன் இராஜாவால் அடிக்கப்பட்டான்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது. இங்குச் செய்வினை வாக்கியம் அடிப்படை வாக்கியமாகவும் (புகைநிலை), செயப்பாட்டு வினை வாக்கியம் வருவிக்கப்பட்ட வாக்கியமாகவும் (புறநிலை) காணப்படுகின்றன. இவற்றிற்கு இடையே காணப்படும் மாற்றத்ததைக் கீழ்வருமாறு காட்டலாம்.

பெ.தொ.1+பெ.தொ.2_ஐ+வினை+காலம்+விகுதிகள்

செ.பா

பெ.தொ.2+பெ.தொ.1-ஆல்

வினை. செய.எ.எச்சம்+பகு+காலம்+விகுதிகள்

என்ற நிலையில் உருவாக்கம் பெறுகின்றது.

5.5.1.2.4 கலப்பு வாக்கியம்

ஒரு வாக்கியத்தின் உள்ளே இன்னொரு வாக்கியத்தை முன்னதன் பகுதியாக அல்லது உறுப்பு வாக்கியமாக இணைக்கும்போது உருவாகும் வாக்கியம் கலப்பு வாக்கியம் எனப்படும். எந்த வாக்கியம் தலைமை வாக்கியம் (Matrix sentence) என்றும், இணைக்கப்படுகின்ற வாக்கியம் உறுப்பு வாக்கியம் (Constituent sentence) என்றும் அழைக்கப்படும். எனவே கலப்பு வாக்கியம் என்பது ஒரு தலைமை வாக்கியத்தையும் ஒன்றோ அல்லது ஒன்றுக்கு மேற்பட்ட உறுப்பு வாக்கியத்தையும் கொண்டிருக்கும் வாக்கியம் ஆகும்.

நான் இராமன் வந்தபோது போனேன்
நான் அங்கு வந்த பையனைப் பார்த்தேன்
போன்றவை கலப்பு வாக்கியங்கள்.

கால வினையெச்சத் தொடர் வாக்கியம்

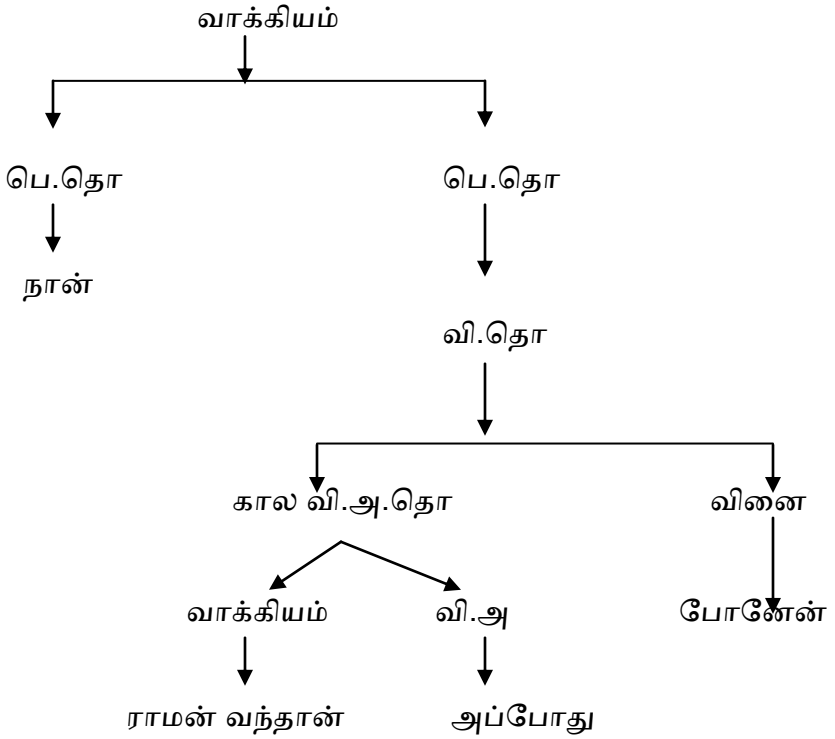
இராமன் வந்தபோது நான் போனேன்
என்பது ஒரு கலப்பு வாக்கியம். இவ்வாக்கியம்

இராமன் வந்தான், அப்போது நான் போனேன்

என்ற இரு வாக்கியங்கள் ஒன்றில் மற்றொன்று உறுப்பு வாக்கியமாக இணைவதால் ஏற்படும் ஒரு வாக்கியம் ஆகும். இதில் இரண்டாவது வாக்கியம் முக்கிய அல்லது தலைமை வாக்கியமாகவும் முதல் வாக்கியம் அதில் கலந்த உறுப்பு வாக்கியமாகவும் உள்ளன. அப்போது என்ற வினையடையை விளக்கும் நிலையில் அமைந்துள்ளது இராமன் வந்தான் என்ற வாக்கியம்.

புதைநிலை (Deep structure)

இவ்வாக்கியத்தின் புதைநிலையைக் கீழ்வருமாறு காட்டலாம்.



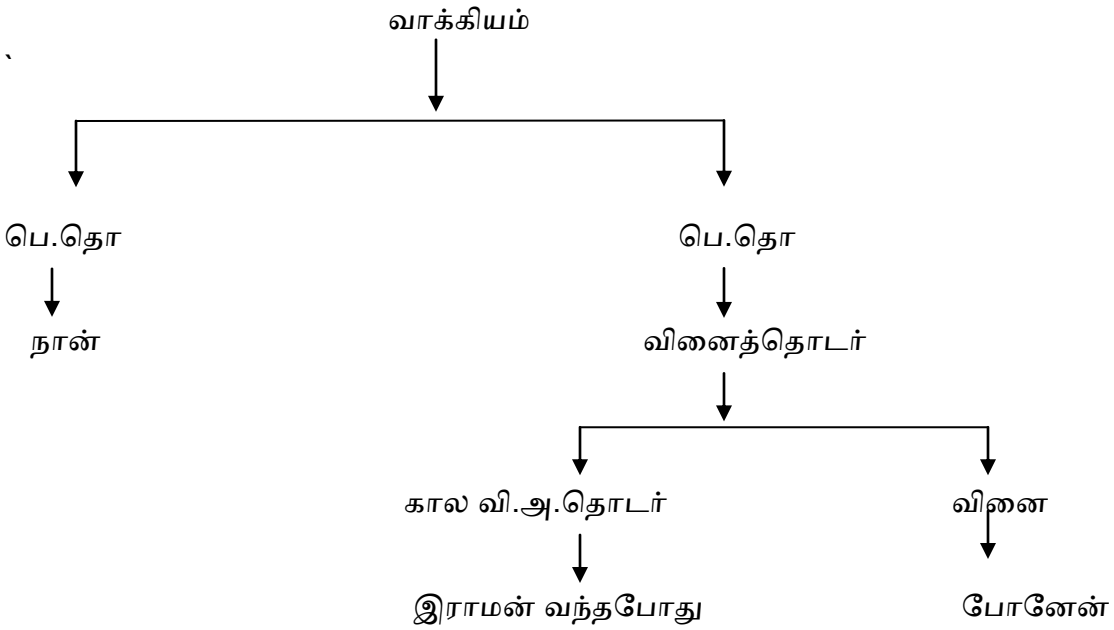
பல்வேறு மாற்றங்கள் (Transformations)

இப்புதைநிலையில் பல்வேறு மாற்றங்கள் நிகழ்கின்றன.

1. வந்தான் என்பது விகுதி. இதுவே பெயரெச்ச விகுதியாகிறது. இதனைப் பெயரெச்ச மாற்றம் (relativisation) எனக் கூறலாம்.
2. அப்போது என்பதில் உள்ள அ என்பது கெடுகின்றது. இதனை அகரச் சுட்டு கெடுதல் எனக் கூறலாம்.

புறநிலை (surface structure)

மேலே காட்டிய மாற்றங்கள் நிகழும்போது நான் இராமன் வந்தபோது போனேன் என்ற புறநிலை வாக்கியம் உருவாகின்றது.



என இதனைக் காட்டலாம்.

பிற கலப்பு வாக்கியங்கள்

பலவிதமான கால எச்சத் தொடர்களும், வேற்றுமைத் தொடர்களும், பின்னுருபுத் தொடர்களும், வினையெச்சத் தொடர்களும், பெயரெச்சத் தொடர்களும் தலைமை வாக்கியத்தின் உறுப்புகளாக வந்து கலப்பு வாக்கியங்களை உருவாக்கும்.

உடனிகழ்ச்சி கலப்பு வாக்கியம்

மணி பாட நான் ஆடினேன்

என்ற வாக்கியத்தில் இரு நிகழ்வுகளைக் குறிக்கும் வாக்கியங்கள் ஒன்றிற்குள் ஒன்றாக உள்ளடக்கப்பட்டு கலப்பு வாக்கியமாக வெளிப்படுகின்றன.

வேற்றுமைத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வேற்றுமைத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

நான் கண்ணன் வந்ததைக் கண்டேன்

பின்னுருபுத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் பின்னுருபுத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நான் வருவதற்கு முன்னர் அவள் வந்து விட்டாள்.

வினையெச்சத் தொடர் கலப்பு வாக்கியம்

இதில் ஒரு வாக்கியம் வினையெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

அவள் சாப்பிட்டுவிட்டு தூங்கினாள்

பெயரெச்சத் தொடர் வாக்கியம்

இதில் ஒரு வாக்கியம் பெயரெச்சத் தொடராக மற்றொன்றில் உட்படுத்தப்பட்டிருக்கும்.

எ.கா

நேற்று ஊருக்குப்போன கண்ணன் இன்று பள்ளிக்கு வரவில்லை.

5.5.1.2.5 கூட்டு வாக்கியம்

இரண்டு வாக்கியங்கள் தலைமை வாக்கியம், உறுப்பி வாக்கியம் என்றில்லாமல் ஒத்த நிலையில் இணையும்போது உருவாகும் வாக்கியம் கூட்டு வாக்கியம் எனப்படும்.

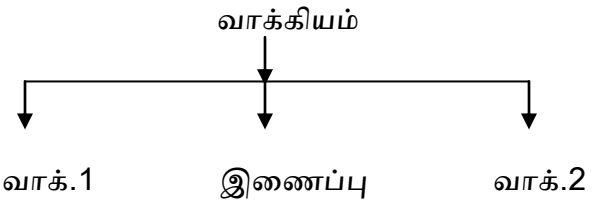
கண்ணன் வந்தான்.

இராமன் வந்தான்.

என்ற இரு வாக்கியங்களும் ஒத்தநிலையில் இணையும்போது

கண்ணனும் இராமனும் வந்தார்கள்.

என்ற கூட்டு வாக்கியம் உருவாகின்றது. இதில் *கண்ணனும் இராமனும்* என்ற பெயர்த்தொடர் எழுவாயாகவும், *வந்தார்கள்* என்பது பயனிலைத் தொடராகவும் உள்ளன. இரண்டு வாக்கியங்களும் கூட்டு வாக்கியத்தில் ஒத்த நிலையில் காணப்படுகின்றன. ஒன்றின் உறுப்பாக மற்றொன்று இல்லை. இத்தகைய இணைப்பு அமைப்பைக் கீழ்க்கண்டவாறு காட்டலாம்.



கூட்டு வாக்கியங்களின் வகைப்பாடு

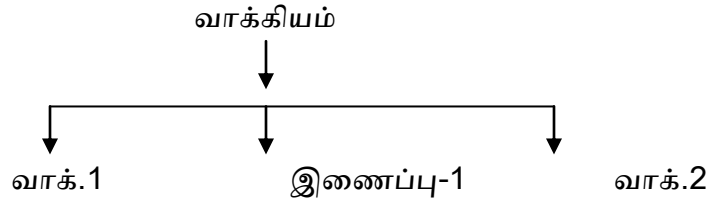
இரு வாக்கியங்கள் இணைந்து உருவாகும் கூட்டு வாக்கியங்கள் (1) முழுநிலை இணைப்பு வாக்கியங்கள் (conjoined sentences) என்றும், (2) பிரிநிலை இணைப்பு வாக்கியங்கள் (disjunctive) என்றும் இரண்டு வகைப்படும்.

முழுநிலை இணைப்பு

முழுநிலை இணைப்பு என்பது ஒன்றுக்கு மேற்பட்ட வாக்கியங்கள் முழு நிலையில் இணைக்கப்படுவது. அதாவது அதுவா இதுவா என்ற நிலையில் இணையாது அதுவும், இதுவும் என்ற கூட்டுநிலையில் இணைக்கப்படுவது (conjunctive).

கண்ணனும் இராமனும் வந்தார்கள்.

இதனைப் பின்வருமாறு காட்டலாம்.



பிரிநிலை இணைப்பு வாக்கியம்

இரு வாக்கியங்களை வேறுபடுத்தி அல்லது பிரித்து இணைக்கும் இணைப்பு பிரிநிலை இணைப்பு (disjunctive) எனப்படும். அதாவது 'அதுவா இல்லை இதுவா' என்ற நிலையில் இணைக்கப்படுவது

கண்ணன் வருவான் அல்லது இராமன் வருவான்.

கண்ணன் அல்லது இராமன் வருவான்.

கண்ணனோ அல்லது இராமனோ வருவார்கள்

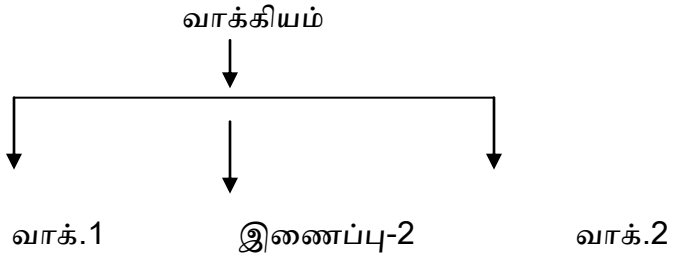
போன்ற வாக்கியங்கள் இப்பிரிவைச் சாரும்.

கண்ணன் வருவான் + இணைப்பு2 + இராமன் வருவான்

என்ற இரு வாக்கியங்களிலிருந்து உருவாவதுதான்.

கண்ணன் வருவான் அல்லது இராமன் வருவான்

போன்ற வாக்கியங்கள் கீழ்வருமாறு காட்டப்படுகிறது.



தமிழ்த்தொடரியல் அமைப்பில் ஒரு பெயர்த் தொடரும் ஒரு வினைத் தொடரும் இணைந்தோ இரு பெயர்த் தொடர்கள் இணைந்தோ ஒரு வாக்கியத்தை உருவாக்கும் வாக்கியங்கள் அவற்றின் அமைப்பின் அடிப்படையில் தனிநிலை வாக்கியம், கூட்டு வாக்கியம் மற்றும் கலவை வாக்கியம் என வேறுபடும். தனிநிலை வாக்கியத்தில் ஒரு வினைத்தொடர் காணப்படும். கூட்டு வாக்கியத்தில் ஒன்றிற்கும் மேற்பட்ட வினைத்தொடர்களோ, பெயர்த்தொடர்களோ இணைப்பான்களால் இணைக்கப்பட்டிருக்கும். கூட்டுநிலை வாக்கியத்தில் ஒன்றுக்கும் மேற்பட்ட வினைத்தொடர்கள் (அதாவது எச்சத் தொடர்கள்) வரும். பொருண்மை அடிப்படையில் வாக்கியங்களைக் கூற்று வாக்கியங்கள், கட்டளை வாக்கியங்கள், வினா வாக்கியங்கள், செய்வினை வாக்கியங்கள், செயப்பாட்டு

வினை வாக்கியங்கள் எனப் பகுக்கலாம். ஒரு வாக்கியம் மற்றொரு வாக்கியத்தில் உட்படையாக வருவதன் காரணமாக வாக்கியங்கள் கலவைத்தன்மை பெறுகின்றன. பெயரெச்சத் தொடர்கள் பெயர்த் தொடருக்குள்ளும் வினையெச்சத் தொடர்கள் வினைத் தொடர்க்குள்ளும் உட்படையாக வரும்.

5.5.2 முந்தைய கணினித் தொடரியல் ஆய்வுகள்

முந்தைய பகுத்துக்குறித்தல் அணுகுமுறைகள் பெரும்பாலும் அகராதியால் இயக்கப்பட்டன. இதில் அகராதி உருபனியல் தரவையும், பொருண்மையியல் வர்ணனைகளையும், தொடரியல் அறிவையும் கொண்டிருக்கும். பின்னர் இலக்கண வடிவ வாதங்களின் தொடக்கம் பல வேறுபட்ட கொள்கைகள் அடிப்படையிலான பகுத்துக் குறிப்பான்களைத் தந்தது. காஷ்கெட் (Kashket,1986) என்பவர் வார்ல்பிரி என்ற மொழிக்கு சாம்ஸ்கிலின் (Chomsky) ஆளகை-கட்டுறவுக் கோட்பாடு (government and binding) அடிப்படையில் பகுத்துக்குறிப்பானை உருவாக்கினார். இவ்வணுகுமுறையில் தொடர்கள், தேர்வுக் கொள்கைகள், வேற்றுமை அடையாளப்படுத்துகை, வேற்றுமை தருகை, பங்கெடுப்பாளர் தொடர்பு படுத்துகை என்பனவற்றின் வழி உருவாக்கப்பட்டன. பங்கெடுப்பாளர் தகவல் அகராதியில் சேகரிக்கப்பட்டன. சென்குப்தா (Sengupta, et al., 1997) போன்றோர் கப்லன் மற்றும் பிரஸ்னைன் என்போரின் சொல் செயல்பாட்டு இலக்கணத்தை (Lexical Fuhnctional Grammar)வங்காள மொழிக்குத் தகுந்தவாறு மாற்றி ஒரு பகுத்துக்குறிப்பானை உருவாக்கினார். சொல் செயல்பாட்டு இலக்கணம் ஆங்கிலம் போன்ற இடம்சார் மொழிகளுக்கு (Positional languages) வலுவான வடிவவாதம் ஆகம். இது உறுப்பமைபபையும் இலக்கண உறவுகளையும் கண்டுகொள்கிறது. செங்குப்தாவின் பகுத்துக் குறிப்பான் செய்வினைப்பாடாக உள்ள எளிய வாக்கியங்களுக்கு நல்ல விளவைத் தந்தது. வினிவார்ட்டர் Winiwarter,1995)ஒன்றிணைப்பு இலக்கணத்தை (unofication grammar) விரிவாக்கி ஜொமன் மொழிக்குப் பயன்படுத்தினார். சினிவார்ட்டரின் ஒழுங்கமைப்பில் இலக்கண விதிகள் அருவமாக்கலின் பொருத்தமான மட்டத்தில் இலக்கண விதிக்ள

அதராதிக்கு பங்கெடுப்பாளராகச் சேர்ககப்ப்பட்டன. பெக்கா போன்றோர் (Becker et.al, 1994) V_TAG என்ற கிளை இணைப்பு இலக்கணத்திலிருந்து (TREE adjoining) மாறுபட்ட ஒன்றான பலபெயர் கால பகுத்துக்குறித்தல் வழிமுறை வரைவை உருவாக்கினார் இப்பகுத்துக்குறிப்பானை கிளை இணைப்பு இலக்கண வடிவவாதம் முற்றிலும் உருவாக்கப்பட்ட மொழிகளுக்கு மட்டுமே பயன்படுத்த இயலும்.

பல பகுத்துக்குறிப்பான்கள் சார்புக் கொள்கைகள் (dependency principles) அடிப்படையில் உருவாக்ககப்பட்டன. கவிங்டன் (Covington, 2003) உரு சுற்று வரைபடத்தால் இயக்கப்பட்ட சொற்களுக்கு இடையே உள்ள சார்புகளை உருவாக்கி வாக்கியங்களுக்கு அமைப்பைத் தரும் ஒரு பகுத்துக்குறிப்பானை உருவாக்கினார். இதில் ஒவ்வொரு வில்லும் சொற்களின் சார்புநிலை உறவை உருப்படுத்தம் செய்யும். காரக உறவுகள் வேற்றுமைக் குறிகளுடன் பொருத்தப்படும் பாணினி இலக்கண அடிப்படையில் பாரதி போன்றோர் Gharti, et, al,1993, 2003) கட்டுப்பாடு அடிப்படையிலான பகுத்துக் குறிப்பான்களின் ஒரு குழுமத்தை உருவாக்கினார். பாரதி ஒழுங்குமுறை பகுத்துக்குறிக்கும் சிக்கலை முழுஎண் வழியமைப்புச் சிக்கலாக மொழி பெயர்த்துத் தீர்வு காண்கிறது. மூர்த்தி (Murthy,1997) இந்திய மொழிகளுக்கு உரக எச்சத்தொடர் அமைப்பு இலக்கணம் (Universal Clause Structure Grammar) என்ற இலக்கண வடிவ வாதத்தை உருவாக்கினார். இவ்வடிவவாதத்தில் மூன்று மட்ட பகுப்பாய்வு பயன்படுத்தப்படுகிறது. தொடர் அடையாளம் காணல், எச்சத்தொடர் அமைப்பு நிர்ணயம் மற்றும் பங்களிப்பு தருதல். ஒவ்வொரு மட்டத்திற்கும் அதன் பகுப்பாய்வுக்குத் தொடர்புள்ள விதிகள் இருக்கின்றன. உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தில் இத்தொகுதியாக்கம் (modularization) இதைப் பிற மொழிகளுக்கு எளிதாகப் பயன்படுத்த செய்தது. சொல்நிரல் சிக்கலைக் கையாளுவதற்கான ஒரு தீர்மாங்களில் ஒன்று வரிசை மாற்று இலக்கணமாகும் (Permutational Grammar (PG)). (Geg. Ikifssib, et, al., 2001) வரிசை மாற்று இலக்கணம் அடிப்படை தொடரமைப்பு விதிகளை அவற்றின் செயல்பாட்டு

உருப்படுத்தங்களுடன் குறிப்பிட்டு பின்னர் ஒரே பொருண்மையைத் தரும் பிற எல்லாத் தொடர்ச்சிகளையும் பெறவேண்டி உறுப்புகளை வரிசைமாற்றம் செய்யும். குமாரசண்முகம் மேலே குறிப்பிட்ட வேறுபட்ட அணுகுமுறைகளின் ஒரு ஒன்றிணைப்பைப் பயன்படுத்துகின்றார். அவர் தொடர்களையும் வாக்கியங்களையும் எச்சத்தொடர்களையும் கண்டுபிடிக்க மூர்த்தியின் உலக எச்சத்தொடர் அமைப்பு இலக்கணத்தைப் போன்று நுட்பங்களைப் (Techniques) பயனபடுத்துகின்றார். சுதந்திர சொல்நிரல் சிக்கலைக் கையாள வரிசைமாற்று இலக்கணத்தைப் போன்று உறுப்புகளை வரிசை மாற்றுவதில்லை. அதற்குப் பதிலாக வாக்கியங்களின் உறுப்புகளை ஒரு குறிப்பிட்ட நிரலில் வரிசைமாற்றம் செய்கிறார். பாரதியின் ஒழுங்கமைப்பில் (Bharati, 2000) உள்ளது போன்று வாக்கியத்தின் எழுவாயையும் செயப்படுபொருளையும் கண்டுபிடிக்க வேற்றுமை உருபுகளைப் பயன்படுத்துகிறார். தமிழின் தொடரியலை விளக்க பொதுவான தொடரமைப்பு இலக்கணததைப் பயன்படுத்துகிறார்.

தமிழ் வாக்கியங்களைப் பகுத்துக் குறிப்பதற்கு வேண்டி ஒரு கணினியியல்சார் சட்டகத்தை உருவாக்கி பயன்படுத்துவதாகும். இதற்கு வேண்டி தமிழ்த்தொடரியலை உருப்படுத்தம் செய்ய தொடரமைப்பு விதிகள் உருவாக்கப்பட்டுள்ளன. நாற்பது மிகக் கூடதலாக நிகழும் வினைகளுக்குப் பங்கெடுப்பாளர் அமைப்பு கண்டுபிடிக்கப் பட்டுள்ளது. பகுத்துக்குறிக்கப்பட்ட வாக்கியங்களின் தொடரியல் அமைப்புகள் பகுத்துக்குறிக்கப்பட்ட கிளைகளில் உருப்படுத்தம் செய்யப்பட்டுளளன. இவ்வடிவில் அவற்றை இயந்திர மொழிபெயர்ப்பு மற்றும் தகவல் பிரித்தெடுப்பு இவற்றை உள்ளடக்கிய பயன்பாடுகளில் நேரடியாகப் பயன்படுத்த இயலும்.

உள்ளீட்டுத் தமிழ் வாக்கியத்தை முக்கியப் பகுத்துக்குறிப்பான் பகுத்துக் குறிப்பதற்கு முன் அது பகுப்பாய்வின் பல மட்டங்களைக் கடந்துசெல்ல வேண்டும். கீழே தந்துள்ள படத்தில் செவ்வகங்கள் பகுப்பாய்வு செய்யும் பகுப்பாய்வுத் தொகுதிகளையும் (analyzer modules) வட்டங்கள் அவை பயன்படுத்தும் பொருத்தமான மூலப்பொருள்களையும் உருப்படுத்தம் செய்யும்.

தமிழுக்கு வேண்டி சில பெயர்த்தொடர் பகுப்பான்கள் உருவாக்கப்பட்டுள்ளன. அவை கீழே விளக்கப்பட்டுள்ளன.

5.5.2.1 AUKBCRC -இன் தமிழுக்கான பெயர்த்தொடர் பகுப்பான்

இவ்வணுகுமுறை விதி அடிப்படையிலானது. இந்நெறிமுறையில் ஒரு தரவுத்தொகுதி எடுக்கப்பட்டு இரண்டோ அதற்கு அதிகமான குமங்களாகப் பிரிக்கப்படும். இப்பிரிக்கப்பட்ட குழுமத்தில் ஒன்று பயிற்சித் தரவாகப் பயன்படுத்தப்படும். பயிற்சித் தரவுக் குழுமம் எடுக்கப்பட்டு பெயர்த்தொடருக்காக மனித முயற்சியால் பகுக்கப்படும். இவ்வாறு ஒரு வாக்கியத்தில் உள்ள பெயர்த் தொடர்களைப் பிரிப்பதற்குப் பயன்படுத்தப்படும் விதிகள் உருவாக்கப்படும். இவ்விதிகள் தொடர் பகுப்பிற்கு அடிப்படையாக அமையும். தொடர் பகுப்பான் வழியமைப்பு இவ்விதிகளைப் பயன்படுத்தி பரிசோதனைத் தரவை பகுக்கும். இவ்விதிகளின் பயன்பாட்டுப் பரப்பு இத்தரவுக்குழுமத்தில் பரிசோதிக்கப்படும். துல்லியமும் மீள அழைத்தலும் இதற்குக் கணிக்கப்படும். விளைவு ஒழுங்குமுறையின் பயன்பாட்டுப் பரப்பை விருத்தி செய்ய கூடதல் விதிகள் தேவையா என்று பரிசோதிக்கப்படும். பின்னர் ஒழுங்குமுறை வேறுபட்ட பிற பயன்பாடுகளுக்குப் பரிசோதிக்கப்படும்.

5.5.2.2 RCILTS –இன் வானவில்

வானவில் இந்திய மொழிகளின் தொழில்நுட்ப முன்னேற்றம் என்ற திட்டத்தின் கீழ் தமிழுக்காகச் செய்யப்பட்ட இயற்கைமொழி ஆய்வுகளில் ஒன்றாகும். இது ஒரு தமிழ் வாக்கியத்தின் தொடரியல் உறுப்புகளை அடையாளம் காண்கிறது. இது பட்டியல் வடிவில் பகுப்புக் கிளைப் படத்தை (parse tree) வெளிப்யிடுகின்றது. இது தனிநிலை வாக்கியங்களைய கூட்டு வாக்கியங்களையும் கையாளுகின்றது. தனிநிலை வாக்கியங்கள் ஒரு வினையையும் பல பெயர்த்

தொடர்களையும், தனிநிலை வினையடைகளையும், பெயரடைகளையும் கொண்டிருக்கும். பல எச்சத் தொடர்கள் உள்ள வாக்கியங்களின் தேர்வுகளில் வானவில் துப்புச்சொற்கள் மற்றும் தொடர்கள் அடிப்படையில் எச்சத் தொடர்களைத் தொடர் குழுக்களாகக் குழுமும். இது தொடரியல் அமைப்பு இலக்கணத்தை உருவாக்கும். வானவில் சொல்நிரலைக் கையாள முன்நோக்குதலைப் பயன்படுத்தும்.

பகுத்துக்குறித்தல் என்பது ஒரு வாக்கியத்தை அறிந்துகொண்டு அதற்குத் தொடரியல் அமைப்பைத் தருவதாகும். இது தொடரியல் ஆய்வில் மிகப்பெரிய சிக்கலாகும். இதன் காரணமாக இது இயற்கை மொழி ஆய்வின் ஒரு பெரும்பகுதியாக அமைகிறது.

நிர்ணயிக்கப்பட்ட பகுத்துக்குறிப்பானின் உள்ளீடு (input) அடையாளப்படுத்தப்பட்ட சொற்களின் நிரலான தொடர்ச்சியும் தொடரமைப்பு விதிகளின் ஒரு குழுமமும் ஆகும். இதன் வெளிப்யீடு (out put) வாக்கியத்தின் பகுத்துக்குறிக்கப்பட்ட அமைப்பாகும். ஒரு வாக்கியத்தின் பகுத்துக் குறிக்கப்பட்ட அமைப்பு அதன் எழுவாய், செயப்படுபொருள் மற்றும் பயனிலைகளை அறிந்துகொள்ள உதவுகிறது. மேலும், பகுத்துக் குறிக்கப்பட்ட கிளை அமைப்பு பிற தொடர்களையும் அடை செய்யும் தொடர்களையும் அடையாளம் காணப்படுகிறது.

ஒரு மொழிக்கு ஒரு மாதிரி பகுத்துக்குறிப்பான் ஒரு அடையாளப்படுத்தியைப் பயன்படுத்தி வாக்கியங்களின் சொல் வகையை அடையாளப் படுத்துகிறது. பின்னர் பகுத்துக்குறித்தும் வழிமுறை வரைவு (algorithm) வாக்கியத்தைப் பகுத்துக்குறித்து அம்மொழி இலக்கணத்திற்குத் தரப்பட்ட தொடரமைப்பு விதிகிளின் அடிப்படையில் பொருத்தமான அமைப்பைத் தரும். பகத்துக்குறித்தலின் போது வழிமுறை வரைவு உறுப்புகளுக்கு இடையேயுள்ள இலக்கண உடன்பாட்டையும் முக்கிய வினைக்கம்

செயப்படுபொருள் தொடர்களுக்கும் இடையே உள்ள சார்பு உறவையும் பரிசோதிக்கும்.

5.5.2.3 குமாரசண்முகத்தின் தொடரியல் பகுப்பாய்வு

குமார சண்முகம் தமது எம்.எஸ். பட்டப்படிப்புக்காக தமிழில் தொடரியல் பகுப்பாய்வு என்ற ஆய்வேட்டைச் சமர்ப்பித்தார். அதன் சுருக்கம் இங்கு தரப்பட்டுள்ளது.

5.5.2.3.1 வாக்கியங்களின் இலக்கணப் பண்புக்கூறுகள்

வாக்கியத்தில் உள்ள சொற்களின் சொல் வகைப்பாடும் பண்புக்கூறு தகவலும் வாக்கியத்தின் பகுத்துக் குறித்தலுக்கு முக்கியமாகும். பின்வரும் அட்டவணை சொல் வகைப்பாடுகளையும் அவற்றுடன் தொடர்புடைய அடையாளங்களையும் அவற்றின் பண்புக்கூறுகளுடன் காட்டும்.

சொல் வகைப்பாட்டுடன் தொடர்புடைய அடையாளப்படுத்திகள்

வ.எண்	சொல் வகைப்பாடு	அடையாளம்/உருப்படுத்தம்	எடுத்துக்காட்டு
1.	முற்றுவினை	V	பார்த்தான்
2.	பெயர்	N	பையன்
3.	மாற்றுப்பெயர்	P	அவர்
4.	பெயரடை	J	நல்ல

5.	வினையடை	R	வேகமாக
6.	பின்னுருபு	I	முன்னால்
7.	சுட்டடை	D	அந்த
8.	அளவடை	G	பல
9.	பெயர் மாறுபடுத்தி	M	புதுமை
10.	கிழமைப்பெயர்	GN	பொருளின்
11.	முற்றுப்பெறாத வினை	NV	பெற்று
12.	பெயரெச்ச வினை	RV	படித்த
13.	நிரப்பிய வினை	CV	என்று
14.	செயல்பாட்டுச் சொல்	F	வரை

5.5.2.3.2 இலக்கணப் பண்புகளும் அவற்றுடன் தொடர்புடைய அடையாளப்படுத்திகளும்

வ.எண்	பண்புகள்	சொல்வரை/உருப்படுத்தம்
பால்		
1.	ஆண்பால்	+M, -F
2.	பெண்பால்	-M, +F
3.	பலர்பால்	+M, +F
4.	ஒன்றன் பால்	-M, -F
எண்		
5.	ஒருமை	+S

6.	பன்மை	-S
இடம்		
7.	தன்மை	+1, -2
8.	முன்னிலை	-1, +2
9.	படர்க்கை	+1, +2
காலம்		
10.	நிகழ்காலம்	+P, -A
11.	இறந்தகாலம்	-P, +A
12.	எதிர்காலம்	+P, +A
வேற்றுமை		
13.	செயப்படுபொருள் வேற்றுமை	+ACC
14.	கொடை வேற்றுமை	+DAT
15.	இட வேற்றுமை	+LOC
16.	கருவி வேற்றுமை	+INS
17.	உடனிகழ்ச்சி வேற்றுமை	+SOC
18.	நீங்கல் வேற்றுமை	+ABL
19.	பயன் வேற்றுமை	+BEN
20.	கிழமை வேற்றுமை	+GEN

5.4.3.3 உருபனியல் பகுப்பாய்வு

உருபனியல் பகுப்பாய்வு சொற்களைப் பகுக்கும் செயல்பாட்டையும் அவற்றின் ஒட்டு உருபன்பளையும் வேர்களைக் கண்டுபிடிக்கும்

செயல்பாட்டையும் உள்ளடக்கும். இவ்வுருபனியல் பகுப்பாய்வினால் ஒரு சொல்லின் சொல் வகைப்பாட்டையும் அதன் இலக்கணப் பண்புக்கூறுகளையும் அடையாளம் காண இயலும்.

5.4.3.4 சொல் வடிவமும் அதன் அடையாளங்களும்

சொல்வடிவம்	உருபன் விடுவரல்	சொல்லும் அதன் அடையாளப்படுத்தியும்
கேட்டான்	கேள்<verb>+த்த்<past tense>+ஆன்<3rd person, singular,masculine>	கேட்டான் <V:+p,-A, +1.+2 +3, +S, +M, -F>
மரத்தை	மரம் <noun> + ஐ <accusative case>	மரத்தை <N:+ACC, +S, -M, -F>

5.4.3.5 சொல்வகை அடையாளப்படுத்தல்

சொல்வகை அடையாளப்படுத்தல் வாக்கியங்களின் உறுப்புகளுக்குச் சரியான அடையாளத்தைத் தரும் செயல்பாடாகும். தமிழ் போன்ற மொழிகளுக்கு உருபனியல் பகுப்பாய்வி 70% துல்லியத்துடன் சொல்வகைப்பாட்டை கண்டுபிடிக்க இயலும். இருப்பினும் பல நிகழ்வுகளில் உருபனியல் பகுப்பாய்வி ஒரு சொல்லுக்கு ஒரு அடையாளத்தை மயக்கமின்றி தர இயலாது. இம்மாதிரியான மயக்கங்களை நீக்கக் குமாரசண்முகம் உருபனியல் பகுப்பாய்விற்குப் பின்னர் சொல்வகை அடையாளப்படுத்தியைப் பயன்படுத்துகிறார். சொல்வகை அடையாளப்படுத்தி உருபனியல் ஆய்வியின் விடுவரல் (output) ஏற்று இயலுகின்ற அளவு சொல்மட்ட விதிகளைப் பயன்படுத்தி மயக்கத்தை நீக்குகிறது. எஞ்சுவது பகுத்துக்குறிப்பானால் கவனித்துக் கொள்ளப்படும்.

5.4.3.6 இயல்பாக்கம் (Normalization)

தமிழ் ஒத்தறி அடிப்படையில் சொல்நிரல் சுதந்திரமுள்ள மொழியாகையால் இறுதி வினைக்கு முன்னர் வரும் பெ.தொ பங்கெடுப்பாளர்கள் இறுதி வினைக்கு முன் எந்த வரிசையிலும் வந்து ஒரே பொருளைத் தரும்.

எ.கா

ஆசிரியர் குழந்தைக்குப் பரிசைக் கொடுத்தார்.

ஆசிரியர் பரிசைக் குழந்தைக்குக் கொடுத்தார்.

பரிசை ஆசிரியர் குழந்தைக்குக் கொடுத்தார்.

பரிசைக் குழந்தைக்கு ஆசிரியர் கொடுத்தார்.

குழந்தைக்குப் பரிசை ஆசிரியர் கொடுத்தார்.

குழந்தைக்கு ஆசிரியர் பரிசைக் கொடுத்தார்.

குமாரசண்முகம் இயல்பாக்கத்திற்காக வாக்கியங்களில் தொடர்கள் ஒரு அனுமதிக்கப்பட்ட எல்லைக்குள் ஏற்கெனவே தீர்மானிக்கப்பட்ட வரிசைகளில் வருவதாக எடுத்துக் கொள்கிறார். இயல்பாக்கி வாக்கியத்தை வரிசையாக ஆய்ந்து தொடர் எல்லைகளை அடையாளம் காண்கின்றது. அதைத் தொடர்ந்து அது எச்சத் தொடர்களையும், பெயர்த்தொடர்களையும் இறுதி வினைக்கு முன் ஒரு குறிப்பிட்ட நிலையில் வரிசைப்படுத்துகிறது. பெயர்த்தொடர்களின் மற்றும் எச்சத் தொடர்களின் எல்லைகள் கண்டுபிடிக்கப்பட்ட பின் இயல்பாக்கி இவற்றை வலமிருந்து இடமாக பின்வரும் முன்நிர்ணயிக்கப்பட்ட நிரலில் மறுவரிசைப்படுத்தும்.

0. வினை,முற்றுப்பெறா வினை,பெயரெச்ச வினை,நிரப்பு வினை

1. பெ.தொ+செயப்படுபொருள் வேற்றுமை

2. பெ.தொ+கொடை வேற்றுமை

3. பின்னுருபு+இட வேற்றுமை

4. பின்னுருபு+கருவி வேற்றுமை

5. பின்னுருபு+உடனிகழ்ச்சி வேற்றுமை

6. பின்னுருபு+நீங்கல் வேற்றுமை

7. பின்னுருபு+பயன் வேற்றுமை

8. வினையடை+கிழமை வேற்றுமை

9. எழுவாய்

எடுத்துக்காட்டாக முன்னர் பட்டியலிடப்பட்ட வாக்கியங்களில் எது தரப்பட்டாலும் இயல்பாக்கி பின்வரும் வெளிப்யீட்டைத் தரும்.

ஆசிரியர் குழந்தைகளுக்குப் பரிசைக் கொடுத்தார்.

இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர் உறுப்பு வரையறுதல்

இந்தியமொழியிலிருந்து இந்திய மொழி மொழிபெயர்ப்பு ஒழுங்கு முறையில் தொடர் உறுப்பு வரையறுதல் ஒருவாக்கியத்தின் தனிநிலை பெயர்தொடர்கள், வினைக் குழுமங்கள், பெயரடைத் தொடர்கள், வினையடைத் தொடர்கள் இவற்றை அடையாளங்காணுதலைக் குறிப்பிடும். இது தொடர் உறுப்புகளின் எல்லையையும் புலக்குறிப்பையும் கண்டுபிடிப்பதை உட்படுத்தும்.

இடுவரல்-விடுவரல் தனிக்குறிப்பீடுகள்

Input-Output Specification

Input: TKN_, CAT_

Output: ADDR_ (ADDR என்பதன் மதிப்பு மாறுகையில் மீளமைப்புச் செயல்பாடு நடக்கிறது என்பது இதன் அர்த்தமாகும்.)

மேலே தரப்பட்டுள்ள இடுவரல் தனிக்குறிப்பீடுகள் SSF-இல் வருணனை செய்வதற்கு TKN_ மற்றும் CAT_ என்ற தனிப்பண்புகள் அதற்குத் தேவை அல்லது அது பன்படுத்துகின்றது என்று அர்த்தம்தரும். வேறுவிதமாகக்

கூறினால் இடுவரல் வாக்கியத்தில் சொற்கள் அல்லது டொக்கன்கள் சொல்வகைப்பாட்டு வகையுடன் தரப்பட்டுள்ளன.

மேலே தரப்பட்டுள்ள விடுவரல் தனிக்குறிப்பீடுகள் 'ADDR_' இந்தத் தொகுதியால் வரையறவிளக்கம் பெறும் என்று அர்த்தம் பெறும்.

எடுத்துக்காட்டு:

வருணனை: தனிநிலைப் பெயர்த்தொடர்கள் தொடர் உறுப்புகளாகக் (NPs) குழுமப்படும். வினைத் தொடர்ச்சிகள் வினைத் தொடர் உறுப்புகளாகக் (VG) குழுமப்படும்.

5.6.1 இ-இ இயந்திர மொழிபெயர்ப்புத் திட்டத்திற்குத் தேர்ந்தெடுக்கப்பட்டுள்ள தொடர் உறுப்புகள்

இதில் தொடர் உறுப்புகளின் வரையறை விளக்கமும் தொடர் உறுப்பு வகைகளும் சில தனித்துவம் வாய்ந்த நேர்வுகளும் பேசப்படும். இந்திய மொழிகளுக்கு இடையிலான மொழிபெயர்ப்புத் திட்டத்தில் 11 தொடர் உறுப்புகள் தெரிந்தெடுக்கப்படு அவை தொடர் உறுப்பு அடையாளப் படுத்தலுக்குப் பயன் படுத்தபடுகின்றன.

முக்கிய தொடர் உறுப்புக் குழுமம் அவற்றின் உருப்படுத்தமும்

	முக்கிய தொடர் உறுப்புக் குழுமம்	**உருப்படுத்தம்**
1	பெயர்த் தொடர் உறுப்பு	NP
2	வினைமுற்றுத் தொடர் உறுப்பு	VGF
3	வினைமுற்றாத் தொடர் உறுப்பு	VGNF
4	வினைப்பெயர்த் தொடர் உறுப்பு	VGNN
5	வினை எச்சத்தொடர் உறுப்பு	VGINF
6	பெயரைடைத் தொடர் உறுப்பு	JJP

7	வினையடைத் தொடர் உறுப்பு	RBP
8	இணைப்புத் தொடர் உறுப்பு	CCP
9	வியப்புத் தொடர் உறுப்பு	INTJ
10	எதிர்மறைத் தொடர் உறுப்பு	NEGP
11	வெற்றிடத் தொடர் உறுப்பு	BLK

5.6.1.1 பெயர்த் தொடர் உறுப்பு (NP)

பெயர்த்தொடர் உறுப்புகள் (NP chunks) NP என அடையாளப் படுத்தப்படும்; இது மறுதரவாக வராத பெயர்த்தொடர்களையும் பின்னுருபுத் தொடர்களையும் உட்படுத்தும். பெயர்த்தொடர் உறுப்பின் தலைமை ஒரு பெயராகும். சிறப்புசெய்யும் சொற்கள் பெயர்த்தொடர் உறுப்பின் இடது பக்க எல்லையை நிறுவும்; வேற்றுமையோ தலைமைப் பெயரோ வலது பக்க எல்லையை நிறுவும். பெயரை அடைசெய்யும் வருணனைப் பெயரடை(கள்) பெயர்த்தொடர் உறுப்பின் பகுதியாக அமையும். தலைமைப் பெயருடன் நங்கூரமிட்டுள்ள குறைச் சொற்களும் பெயர்த்தொடர் உறுப்புடன் குழுமப்படும். அது பெயருக்குப் பின்னரோ வேற்றுமைக்குப் பின்னரோ வந்தால், அது தொடர் உறுப்பின் வலது பக்க எல்லையை நிர்ணயிக்கும். பெயர்த்தொடர் உறுப்புகளின் சில எடுத்துக்காட்டுகள் கீழே தரப்பட்டுள்ளன.

((குழந்தைகள்_NN))_NP, ((சில_QF குழந்தைகள்))_NP, ((சில_QF நல்ல_JJ குழந்தைகள்_NN))_NP, ((மேசை மேல்_PSP))_NP, ((ஒரு_QC கறுத்த_JJ குதிரை_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகம்_NN))_NP, ((இந்த_DEM புதிய_JJ புத்தகத்தின்_NN மேல்_PSP))_NP

5.6.1.2 வினைத் தொடர் உறுப்பு (VG)

வினைத் தொடர் உறுப்புகள் பலவகைப்படும். ஒரு வினைக் குழுமம் முக்கிய வினையையும் இருப்பின் துணை வினைகளையும் உள்ளடக்கும். பின் வருவது வினைத் தொடர் உறுப்புகளின் எடுத்துக்காட்டுகளாகும்.

((போனான்)), ((போய்விட்டான்)), (போய்க்கொண்டிருக்கிறான்))

5.6.1.3 பெயரைடைத் தொடர் உறுப்பு (JJP)

பெயரடைத் தொடர் உறுப்பு JJP என அடையாளப்படுத்தப்படும். இத்தொடர் பயனைலைப் பெயர்தொடர்களையும் உள்ளடக்கி எல்லாப் பெயர்த்தொடர் உறுப்புகளையும் கொண்டிருக்கும். இருப்பினும் பெயருக்கு முன்னர் வரும் பெயரடைகள் பெயர்த் தொடர் உறுப்புடன் சேர்த்துக் குழுமப்படும்.

5.6.1..4 வினையடைத் தொடர் உறுப்பு (RBP)

இத்தொடர் உறுப்பின் பெயரும் சொல்வகை அடையாளப் படுத்தலின் போது பயன்படுத்தப்படுகிற சொல்வகை அடையாளத்தின் அடிப்படையில் அமைகின்றது. எல்லா வினையடைத்தொடர் உறுப்பும் RBP என அடையாளப்படுத்தப்படும். இந்தத் தொடர் உறுப்பு எல்லாச் சுத்தமான வினைத்தொடர்களையும் உட்படுத்தும்.

5.6.1.5 இணைப்புத் தொடர் உறுப்பு (CCP)

இணைப்புத் தொடர் உறுப்புகள் யாவும் CCP என அடையாளப்படுத்தப்படும். இத்தகவல் பெரிய வாக்கிய அமைபுகளை உருவாக்கத் தேவையாகும்.

5.6.1.6 வியப்புத் தொடர் உறுப்பு (INTJ)

எல்லா வியப்புத் தொடர் உறுப்புகளும் INTJ என அடையாளப்படுத்தப்படும்.

5.6.1.7 எதிர்மறைத் தொடர் உறுப்பு (NEGP)

வினையைச் சுற்றி இருக்கும் எதிர்மறைக் குறைச்சொல் வினைக் குழுவுடன் குழுமப்படவேண்டும். அது வினையிலிருந்தோ பெயரிலிருந்தோ விலகி இருந்தால் அது தனியான தொடர் உறுப்பாக குழுமப்படவேண்டும்.

6.5.2. தொடர் உறுப்பு வரையறுத்தலின் எடுதுக்காட்டுகள்

AU-KBC-நிறுவனதில் நடைபெற்றுவரும் இந்திய மொழியிலிருந்து இந்திய மொழிபெயப்பு (தமிழ்-இந்தி) ஒழுங்குமுறையின் இரு கூறுகள் இவ்வாய்விற்கு எடுத்துக்கொள்ளப்பட்டுளன என்பதன் அடிப்படையில் இவ்விரு செயல்பாடுகளையும் வெளிப்படுத்தும் பொருட்டுத் தமிழ் தரவுத் தொகுதியில் எவ்வாறு சொல்வகை அடையாளப்படுத்தலும் தொடர் உறுப்பு பிரித்தலும் நடைபெறுகின்றன என்பதைக் கூறும் முகமாகக் கீழே எடுத்துக்காட்டு தரப்பட்டுள்ளது.

<NP> தலம்/NN :/SYM </NP> <NP> வைணவ/JJ தலம்/NN </NP>

<NP> அமைவிடம்/NN :/SYM </NP> <NP> திருச்சியிலிருந்து/NN </NP> <NP>

3/QC மைல்கள்/NN தூரம்/NN </NP>

<NP> திருக்கோயிலின்/NN பெயர்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC

அரங்கநாதசுவாமி/NNP திருக்கோயில்/NN </NP>

<NP> இறைவன்/NN /SYMஇறைவி/NN திருநாமங்கள்/NN :/SYM </NP> <NP> ஸ்ரீ/NNC அரங்கநாத/NNPC சுவாமி/NNP </NP> <NP> ஸ்ரீ/NNC ரங்கநாச்சியார்/NNP </NP>

<NP> தீர்த்தங்கள்/NN :/SYM </NP> <NP> சந்திரபுஷ்கரணி/NN </NP> <CCP> மற்றும்/CC </CCP> <NP> சூரிய/JJ புஷ்கரணி/NN </NP>

<NP> முக்கிய/JJ விழாக்கள்/NN :/SYM </NP> <NP> வைகுண்ட/NNC ஏகாதசி/NN பிரம்மோற்சவம்/NN </NP>

<NP> சிறப்பு/NN அம்சம்/NN :/SYM </NP> <NP> 108/QC வைணவத்/JJ திருத்தலங்களில்/NN </NP> <JJP> முதன்மையான/JJ </JJP> <VGF>திருக்கோயிலாகும்/VM ./SYM </VGF>

<NP> 12/QC ஆழ்வார்களில்/NN </NP> <NP> 11/QC ஆழ்வார்கள்/NN </NP> <NP> இதன்/PRPபுகழை/NN </NP> <NP> பெருமையை/NN </NP> <NP> கீர்த்தியைப்/NN </NP> <VGF>பாடியுள்ளார்கள்/VM ./SYM </VGF>

<NP> இந்த/DEMதலத்தில்தான்/NN </NP> <NP> ஆண்டாள்/NNP </NP> <NP> திருப்பாணாழ்வார்/NNP </NP> <NP> பீபிநாச்சியார்/NNP </NP> <CCP> மற்றும்/CC </CCP> <NP> பலர்/NN </NP> <NP> அரங்கனின்/NNP திருவடியை/NN </NP> <VGF>அடைந்தனர்/VM ./SYM </VGF>

<RBP> சுயம்புவாக/RB </RBP> <NP> உருவான/JJ ஒரே/QC திருக்கோயில்/NN ./SYM </NP>

<NP> போக்குவரத்து/NN மையம்/NN :/SYM </NP> <NP> திருச்சி/NNP ./SYM </NP>

<NP> அரங்கன்/NNP </NP> <VGNF> குடிகொண்டுள்ள/VM </VGNF> <NP> திருவரங்கம்/NNP </NP<NP> திருச்சி/NNP /SYM</NP> <NP> சென்னை/NNP கார்டு/NN ரெயில்/NN பாதையில்/NN </NP> <BLK> உள்ள/PSP </BLK> <NP> ரேVயில்/NN </NP> <VGF>நிலையமாகும்/VM ./SYM </VGF>

<NP> சென்னையிலிருந்து/NNP </NP> <NP> 207/QC மைல்கள்/NN தூரத்திலும்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP சந்திப்பிலிருந்து/NN </NP>

<NP> 3/QC மைல்கள்/NN தொலைவிலும்/NN </NP> <VGF>உள்ளது/VM ./SYM </VGF>

<NP> திருவரங்கம்/NNP நகரம்/NN </NP> <NP> திருச்சிராப்பள்ளி/NNP மாநகரின்/NN புறநகரில்/NN </NP> <BLK> உள்ள/PSP </BLK> <JJP> சிறிய/JJ </JJP> <VGF>நகரமாகும்/VM ./SYM </VGF>

<NP> ஒருபுறம்/NN காவிரியும்/NNP </NP> <NP> மற்றொருபுறம்/NN கொள்ளிடமும்/NNP </NP> <VGNF>பாய்ந்தோடும்/VM </VGNF><NP> இந்த/DEMதிருத்தலத்துக்கு/NN </NP> <RBP> ஆண்டுமுழுவதும்/RB </RBP> <NP> ஆயிரக்கணக்கான/JJ இந்துக்கள்/NN </NP> <VGNF>வந்து/VM </VGNF><NP> அரங்கனை/NNP </NP> <VGF>வழிபட்டு/VM செல்கின்றனர்/VAUX ./SYM </VGF

<RBP> குறிப்பாக/RB </RBP> <NP> வைஷ்ணவர்களுக்கு/NN </NP> <NP> இதுPSP</NP> <JJP> முக்கியமான/JJ </JJP> <VGF>இடமாகும்/VM ./SYM </VGF>

இயல் 6

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை (ILTLMT system)

6.1. அறிமுகம்

6.1.1 நோக்கம்

இந்திய மொழிகளிலிருந்து இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்புத் திட்டம் இணையும் அடிப்படையிலான மொழிபெயர்ப்பு திட்டம் 2006 அக்டோபரில் மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில் நுட்ப அமைச்சின் கீழ் தகவல் தொழில் நுட்பத் துறையின் நிதி நல்கையால் நடைபெற்று வருகின்றது. இதன் முதல் நிலை 2006 அக்டோபர் 30-ஆம் நாள் தொடங்கப்பட்டு 2010 ஏப்பிரலில் முடிவுறது. இத்திட்டம் இந்திய மொழியிலிருந்து மற்றொரு இந்திய மொழிக்கு இயந்திர மொழிபெயர்ப்பை உருவாக்கம் இணையம் அடிப்படையிலான மொழிபெயர்ப்புத் திட்டமாகும். இந்த ஒழுங்குமுறை இரு திசையில் செயல்படும் மற்றும் பொதுவான களத்தில் பின்வரும் மொழிகளுக்கிடையில் செயலாற்றும்.

தமிழ் - இந்தி

தெலுங்கு - இந்தி

மராத்தி - இந்தி

பெங்காளி - இந்தி

தமிழ் - தெலுங்கு

உருது - இந்தி

கன்னடா - இந்தி

பஞ்சாபி - இந்தி

மலையாளம் - தமிழ்

இந்திய மொழி பேசுபவர்கள் இந்த ஒழுங்குமுறையைப் பயன்படுத்துவார்கள். இந்த ஒழுங்குமுறையைக் கற்றல் நுட்பங்களைப் (Machine learning teachiques) பயன்படுத்தி மேலும் இயலும். இந்த ஒழுங்குமுறை இரண்டு குறிப்பிட்ட களங்களுக்கு இயந்திர மொழிபெயர்ப்பை செய்யும்.

பின்வரும் அடிப்படைக் கருவிகள் இந்திய மொழி மொழிப்பெயர்ப்பு ஒழுங்குமுறையின் பகுதியாக உருவாக்கப்பட்டுள்ளது.

1. சொல்வகைப்பாட்டு அடையாளப்படுத்தி (POS Tagger)
2. தொடர் பகுப்பான் (Chunker)
3. உருபனியல் பகுப்பாய்வி (Morph Analyser)
4. வட்டாரச் சொல் குழும (Local word Grouper)
5. மாற்றமைவுப் பகுதி (Transfer Module)
6. இருமொழியப் பொருத்தம் (Billingual Mapping)
7. வாக்கிய நிலை உருவாக்கம் (sentence level generation)
8. வட்டாரச் சொல் குழுமப் பிரிப்பான் (local word group splitter)
9. உருபனியல் உருவாக்கம் (morphological generation)

6.1.2 வாய்ப்பு (scope)

இந்த ஒழுங்குமுறை நிலைப்பேறு பெற்ற மொழியில் எழுதப்பட்ட இணைய பக்கங்கள் அல்லது கால இதழ்கள், நாளிதழ்கள் மீது வேலை செய்யும். சரியான தன்மை பயன்படுத்துவோர் திருப்தி (accuracy / user satisfaction) 85 % - 90 % விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிப்பெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்பட்டுள்ளது மற்றும் 80– 85 % திறமையுடன் பொதுப் பயன்பாடு இயந்திர மொழிபெயர்ப்பு உருவாக்கப்படுள்ளது.

6.1.3. சரியான தன்மை/ பயன்படுதுவோர் திருப்தி (Accuracy/user satisfaction)

85-90 விழுக்காடு திறமையுடன் களம் சிறப்பிக்கப்பட்ட இயந்திர மொழிபெயர்ப்பின் முழுமையான அமைப்பு உருவாக்கப்படும். மற்றும் 80-85 விழுக்காடு திறமையுடன் பொதுப்பயன்பாடு இயந்திர மொழிபெயப்பு உருவாக்கப்படும்.

6.1.4 ஒழுங்குமுறைச் சிறப்பு (system specification)

இந்திய மொழிபெயர்ப்புத் திட்டம் இணையத்துடன் தொடர்பு கொண்டு பயன்படுத்துவோருக்குத் தானியக்க மொழிபெயர்ப்பு சேவைஇயைத் தரவேண்டி ஒரு பரிமாறி (Server) இயங்கும். பின்வருவன மென்பொருள் சிறப்பீடுகளாகும்.

Plat form	:	Fedora Core 4 (Kernel – 2.6.11)
Web server	:	Apache – 2.0
Data base	:	GDBM – 1.8
Compiler	:	gcc 4.0

6.1.5 நெறிமுறை (methodology)

இந்திய மொழி இயந்திர மொழிபெயர்ப்பு ஒழுங்கு முறை பல இயற்கைமொழி ஆய்வு ஆய்வுக் குழுவினர் பங்களிப்பால் உருவாக்கப்பட்டுளது. இந்திய மொழி பெயர்ப்பு ஒழுங்குமுறையை உருவாக்கம் பெரிய செயல்பாடு சிறிய செல்பாடுகளாப் பகுக்கப்பட்டுள்ளது. பங்களிக்கும் ஒவ்வொரு குழுவினரும் ஒன்றோ அதற்கு மேலோ வேளைகளை எடுத்துக்கொண்டு இணையாக வேலை செய்துள்ளனர். பெரும்பாலான பகுதிகள் மொழிச் சுதந்திரமான (language independent engine) இயந்திரமாகவும் மொழிச் சிறப்பான தரவாகவும் (language specific data) பகுக்கப்பட்டுள்ளது. இது போல் ஒரு மொழிக் குழுவினர் ஒரு குறிப்பிட்ட மொழிக்குத் தரவைத் தருவர். ஒரே இயந்திரம் பல மொழிகளுக்கு வேலை செய்யும்.

5.1.5.1 பகுதிகளாக அமைத்தல்

இந்திய மொழிபெயர்ப்பு ஒழுங்குமுறை பல பகுதிகளைக் கொண்டது. ஒவ்வொரு பகுதியும் ஒரு தர்க்கம்சார் வேலையைச் செய்யும், பெரும்பாலும் இந்த வேலை சிறியதாகும். இதன் காரணமாக ஏதாவது மாற்றம் ஏற்பட்டால் சிறிய பகுதியில் அதை எளிதாகச் செய்யவியலும்.

5.1.5.2 சக்தியின் தரமான அமைப்பு (Shakthi Standard Format (SSF)

எல்லாப் பாகங்களும் (modules)வடிவ அமைப்பு நிர்ணயிக்கப்பட்ட தரவின் ஒழுக்கில் செயல்படும். ஒரு பகுதி அதன் வேலையில் வெற்றி பெற்றால் அது வெளிPயீடு ஒழுங்குப் புதிய சின்னத்தையோ ஆய்வையோ சேர்க்கும். பெரும்பாலான பகுதிகளுக்குத் தனிப்பகுதிகளின் வட்டாரச் சிக்கல்களை உறுதியான கட்டுப்பாட்டில் வைக்க இந்த அணுகுமுறை உதவும்.

5.1.5.3 தோல்விகளை நேர்செய்தல்

(SSF) வடிவமைப்பு பகுதியின் ஒவ்வொரு நிலையில் ஏற்படும் தோல்விகளை நேர்செய்ய வசதி செய்யும். எடுத்துக்காட்டாக அதே SSF -இன் மீது அதன் சின்னத்தின் மதிப்பை விளைவிக்கத் தவறினால் கீழ் ஒழுங்குப் பகுதிகள் (down streom modules) தரவின் ஒழுங்கின் மீது தொடர்ந்து செயல்படும்.

5.1.5.4 தெளிவு

இந்தியமொழி இயந்திர மொழிப்பெயர்ப்பில் SSF வடிவமைப்பின் பயன்பாடு ஒவ்வொரு பகுதியின் உள்ளீட்டிற்கும் வெளிப்யீட்டிற்கும் முன் நிகழ்ந்திராதத் தெளப்வைப் பெற உதவுகிறது. ஒரு பகுதியின் உரை சார் வெளிப்யீடு மனிதப் பயன்பாட்டிற்கு மட்டுமல்லாமல் அதன் உள்ளீடாக தரவு ஒழுக்கில் அடுத்து வருகிற பகுதியிலும் பயன்படுத்தப்படும்.

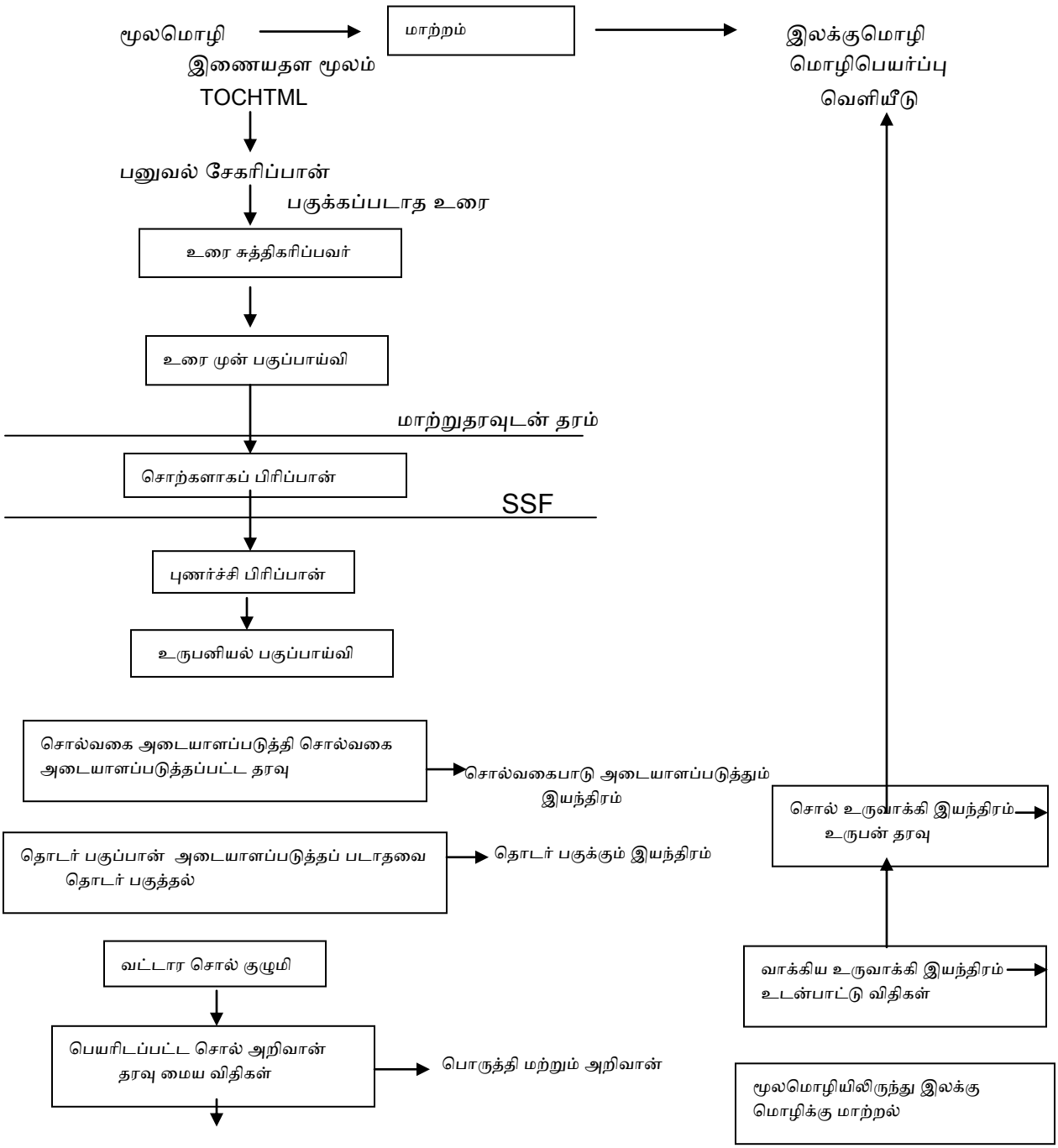
5.1.5.5 டாஷ்போர்டு

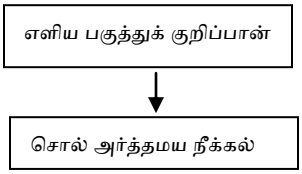
இந்திய மொழி மொழிபெயர்ப்பின் எல்லாப் பகுதிகளும் டாஷ்போர்டின் (dash board) உதவியால் இயங்குகிறது. இது குறிப்பிட்டபடி தரவு ஒழுக்கின் பைப்லைன்களை உருவமைப்பு செய்கிறது. இருப்பினும் பைப் லைன்களுக்குப் பதிலாக எல்லாப் பகுதிகளும் செயல்படும் பங்கிடப்பட்ட நினைவகத்தையோ பிளாக் போர்டையோ (Black Board) நிறுவுவதால் ஒங்குமுறையை விரைவுப்படுத்த இயலும். இருப்பினும் இது பங்கிடப்பட்ட நினைவாகத்தில் (அல்லது நினைவக உருப்படுத்தத்தில் SSF) இயங்குவதற்கு வேண்டி பகுதிகள் எழுதப்பட்டிருக்கவேண்டும் என்று அவசியப்படும்.

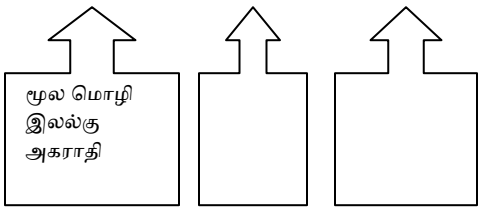
5.2 முழுவிளக்கம்

இந்திய மொழியிலிருந்து இந்திய மொழிக்கான மொழிபெயர்ப்பு ஒழுங்குமுறை ஆய்தல் - மாற்றியமைத்தல் - உருவாக்குதல் என்ற மேற்கோள்

வாய்ப்பாடு அடிப்படையில் செயல்படும். முதலாவது மூலமொழியின் ஆய்வு செய்யப்படும். பின்னர் இலக்குமொழிக்குச் சொற்றொகையும் ஆயப்பட்ட அமைப்பும் மாற்றம் செய்யப்படும். இறுதியாக இலக்கு மொழி உருவாக்கப்படும். இந்திய மொழிகள் ஒற்றுமையுள்ளதாக இருப்பதாலும் இலக்கண அமைப்புகளைப் பங்கீட்டுக் கொள்வதாலும் ஆழமற்ற பகுத்துக் குறித்தல் செய்யப்படும். மாற்றமைவு இலக்கணப் பகுதி எளிமையாக வைக்கப்படும். பொருண்மை களச் சிறப்புப் பெயரிடப்பட்ட சொற்களின் புரிந்து கொள்வான்களாலும் பொருத்தமான அகராதி போன்றவற்றாலும் கையாளப்படும். ஆழமற்ற பகுத்துக் குறித்தலானது, உருபனியல் ஆய்வு, சொல் வகைப்பாடு அடையாளப்படுத்தல் மற்றும் தொடர்பகுத்தல் என்பனவற்றை உள்ளடக்கும். முதலாவது (உருபனியல் பகுத்தாய்வு) விதி அடிப்படையிலானது, இரண்டாவது (சொல் வகைப்பாடு அடையாளப்படுத்தல்) புள்ளியியல் அடிப்படையிலானது, மூன்றாவது (தொடர்ப்பு) விதி அடிப்படை மற்றும் புள்ளியியல் அடிப்படை இரண்டும் இணைந்தது. முழு ஆய்வு வேலையும் பல பாகங்களைப் பிரிக்க்பட்டுள்ளது. ஒவ்வொரு பாகமும் சிறிய தர்க்க வேலையைச் செய்யும். இந்த ஒழுங்குமுறையின் முழு அமைப்பும் கீழே தரப்பட்டிருக்கும் படத்தில் காட்டப்பட்டுள்ளது.


மூலமொழி
மாற்றம்
இலக்குமொழி
இணையதள மூலம்
மொழிபெயர்ப்பு
TOCHTML
வெளியீடு
பனுவல் சேகரிப்பான்
பகுக்கப்படாத உரை
உரை சுத்திகரிப்பவர்
உரை முன் பகுப்பாய்வி
மாற்றுதரவுடன் தரம்
சொற்களாகப் பிரிப்பான்
SSF
புணர்ச்சி பிரிப்பான்
உருபனியல் பகுப்பாய்வி
சொல்வகை அடையாளப்படுத்தி சொல்வகை அடையாளப்படுத்தப்பட்ட தரவு
சொல்வகைபாடு அடையாளப்படுத்தும் இயந்திரம்
சொல் உருவாக்கி இயந்திரம்
உருபன் தரவு
தொடர் பகுப்பான் அடையாளப்படுத்தப் படாதவை தொடர் பகுத்தல்
தொடர் பகுக்கும் இயந்திரம்
வட்டார சொல் குழுமி
வாக்கிய உருவாக்கி இயந்திரம்
உடன்பாட்டு விதிகள்
பெயரிடப்பட்ட சொல் அறிவான் தரவு மைய விதிகள்
பொருத்தி மற்றும் அறிவான்
மூலமொழியிலிருந்து இலக்கு மொழிக்கு மாற்றல்





ஒவ்வொரு முக்கியமான பகுதியும் கீழே விளக்கப்பட்டுள்ளன. ஒவ்வொரு பகுதியும் அதற்கு நேரே தரப்பட்டுள்ள முக்கிய வேலையைச் செய்கின்றன. இருப்பினும் முக்கிய வேலை தவிர அவை சில முன்னாய்வையும் (Preprocetsins) பின்னாய்வையும் (Post processing) செய்கின்றன. எடுத்துக்காட்டாகத் தொடர்பகுப்பான் பகுது தொடர் பகுதிகளைக் கண்டுபிடிக்கின்றன. இவ்வாறு தொடர்பகுப்பான் பகுது முக்கிய வேலையான தொடர் பகுத்தலைச் செய்கின்றது. இருப்பினும் தொடர் பகுதியின் தலைமையைக் கண்டுபிடிக்கும் பின்னாய்வைச் செய்கின்றது. மேலும் தலைமைப் பண்புக் கூறுகளைத் தொடர்பகுதியின் பண்புக்கூறுகளாக நகல் செய்கின்றது.

5.2.1 உருபனியல் பகுப்பாய்வு

ஒரு சொல் தரப்படும் போது உருபனியல் ஆய்வி அதன் வேரையும் சொல்லின் இலக்கணப் பண்புக்கூறுகளையும் அடையாளங்கண்டு கொள்கிறது. திரிபு வள்ம் இல்லாத மொழிகளுக்கு எல்லாச் சொல் வடிவுகளையும் கொண்ட எளிய அகராதி நோக்கீடு போதுமானது. தமிழ், தெலுங்கு போன்ற திரிபு வளமுள்ள மொழிகளுக்கு அகராதியை உருவாக்குவது எளிதல்ல. இது பெரிய சேகரிப்பையும் உயர்ந்த செயல்பாட்டுக் கணிப்பையும் வேண்டும். இதற்கு நல்ல மாற்று வழி வேர்ச்சொற்களின் அகராதியை வைத்துக் கொண்டு உருபனியல் ஆய்வியைப் பயன்படுத்தி சொ வடிவை வேராகவும் இலக்கணப் பின்னுருபுகளாகவும் பிரிப்பதாகும்.

மேற்கோள் வாய்ப்பாடு அடிப்படையிலான அணுகுமுறைகள் (Paradigm based approaches) இந்திய மொழிகளுக்கு நல்ல வெற்றி மதிப்பீட்டைத் தருகிறது.

5.2.2 சொல்வகைப்பாடு அடையாளப்படுத்தி

சொல்வகை அடையாளப்படுத்தல் வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் ஒரு சொல் வகைப்பாட்டைத் தருவதாகும். வாக்கியத்தில் உள்ள ஒவ்வொரு சொல்லுக்கும் பெயர்கள், வினைகள், பெயரடைகள், வினையடைகள் என்று அடையாளம் தருவது ஒரு வாக்கியத்தில் ஒவ்வொரு உறுப்பு பின் ஒரு பங்களிப்பை ஆய்வதற்கு உதவும். சொல்வகைப்பாட்டு அடையாளப் படுத்துவதற்கு விதி அடிப்படை, புள்ளியியல் அடிப்படை போன்ற பல அணுகுமுறைகள் இருக்கின்றது. இங்கு புள்ளியல் அணுகுமுறை பின்பற்றப்படுகின்றது.

5.2.3 தொடர் பகுப்பான்

தொடர்பகுத்தல் (Chunking) ஒரு வாக்கியத்தில் உள்ள எளிய பெயர்த்தொடர்கள், வினைக்குழுமங்கள், பெயரடை தொடர் மற்றும் வினையடை தொடர்களை அடையாளம் காண்பது ஆகும். இது தொடர் பகுது மற்றும் புலக்குறிப்பு இவற்றின் எல்லையை அடையாளங்காண்பதை உட்படுத்தும்.

5.2.4 வட்டாரச் சொல் குழுமி (Local Word Grouper)

வட்டாரச் சொல் குழுமி விபக்தி கணினியக்கத்தின் (Vibhakti Computation) தொழில்நுட்ப வேலையைச் செய்யும் தனிச்சொற்களாக வரும் துணை வினைகள் தனி பொருளடக்கச் சொற்களாகக் கருதப்படுவதில்லை. ஏனென்றால் அவை முதன்மை வினையின் தலைமையாக இருக்கும் தொடர்பகுதியின் இலக்கணப் பண்புக் கூறுகளைச் சிறப்பீடு செய்யும். இது இந்தியில் தனிச் சொற்களாக

வருகின்ற பெயர்விபக்தி என்று அழைக்கப்பெறிகின்ற பின்னுருபுகளுக்கும் உண்மையாகும். வட்டாரச் சொல் குழுமி பொருத்தமான கலவை நிலை மாறும் கூட்டுநிலை சொல் குழும ஆக்க விதிகளைத் திறமையாகக் கையாளும். முக்கியமான வேலை வட்டாரத் தகவல் அடிப்படையில் செயல்பாட்டுச் சொற்களைப் (Functional Words) பொருளடக்கச் சொற்களுடன் (Content Words) குழுமுவதாகும். இந்த வேலையை நிறைவேற்ற பயன்படுத்தப்படும் ஒரு செயல்முறை கிரியரூப்பா அட்டவணைகளைப் (Kiya rupa Charts) பயன்படுத்துவதாகும். இந்த அட்டவணைகள் ஒரு தனி செயல்பாட்டைக் குறிக்கும் தொடர்ச்சியான வினைகளிலிருந்து உருவாக்க வேண்டிய குழுமங்களைச் சிறப்பீடு செய்யும்.

5.2.5 மாற்றம் பகுதி (Transfer Module)

மொழியாய்வில் வாக்கியங்கள் அவற்றின் தொடரமைப்பை அடையாளம் காண்பதற்காகப் பகுத்துக் குறிக்கப்பெறும். இந்திய மொழிகளுக்குள் வேறுபாடுகளை விட ஒற்றுமைகள்தான் அதிகம். எடுத்துக்காட்டாகத் தமிழ், இந்தி மொழியினை முழு பகுத்துக் குறித்தலை வேண்டாது. இத்திட்டத்தில் இயந்திர மொழிபெயர்ப்பு முழு வாக்கியப் பகுத்துக் குறிப்பான் இல்லாது நிறைவேற்றப்படும். மூலமொழியின் அமைப்பிற்கு இலக்குமொழி அமைப்பில் நிகரான அமைப்பு இல்லாவிட்டால் அமைப்பு சார் மாற்றம் தேவைப்படும். ஒரு பகுதி பகுத்துக் குறித்தல் அல்லது ஆழமற்ற பகுத்துக் குறித்தல் மாற்றாத்திற்கு உட்படுத்தப்படும் வாக்கியங்களின் குறிப்பிட்ட உறுப்புகளை அடையாளம் கண்டால் போதுமானது.

மொழிகளுக்கிடையில் உள்ள தொடரியல் வேறுபாடு கலவை வாக்கிய கட்டுமானங்களில், குறிப்பாக எச்சத்தொடர்களில் காணப்படும். இது மாற்றமைவு இலக்கணத்தால் இணைக்கப்படும். இந்தப் பகுது எழுத்துருவாக்கத்தின் வேலையையும் உட்படுத்தும். இந்திய மொழிகளுக்குள் (உருது மொழியை

உள்ளடக்கி) எழுத்துருவாக்கத்தை நிறைவேற்றும் ஒரு பகுதி உருவாக்கப்பட வேண்டும். எழுத்துருவாக்கம் சொல்லையோ சொற்களையோ படிப்பவரின் எழுத்துக்கு மாற்றுகிறது. எடுத்துக்காட்டாக ஒருவருக்கு இந்து தெரிந்திருந்தால் தேவநகரியில் வங்காள உரையைப் படிக்க இயலும் என்றால் அதன் பொருளின் சில பாகங்களை அவரால் புரிந்து கொள்ள முடியும். மொழிபெயர்ப்பு ஒரு சொல்லுக்கோ ஒரு தொடர் பகுதிக்கோ தோற்றுப்போனால் எழுத்துருவாக்கம் படிப்பவரை படிக்கவும் புரிந்து கொள்ளவும் அனுமதிக்கும். இந்திய மொழிகள் பல எண்ணிக்கையிலான சொற்களைப் பங்கிட்டுக் கொண்டுள்ளன. எனவே படிப்பவருக்கு ஏற்றவாறு எழுத்துருமாற்றம் செய்தால் அவரால் மூலமொழியில் உள்ள உரையிலிருந்து சிலவற்றைப் புரிந்து கொள்ள இயலும்.

5.2.6 இருமொழிய அகராதி பொருத்தம் (Bilingual Mapping)

உருபனியல் பகுப்பாய்வால் அடையாளம் காணப்பட்ட வேர்ச்சொல் இலக்குமொழி நிகரன்களுக்கு வேண்டி இருமொழி அகராதியில் பார்க்கப்படும். இந்த இருமொழி அகராதி இலக்கு மொழியின் வேர்ச்சொல் நிகரன்களையும் அதன் வகைப்பாடு (பெயர், வினை போன்றவை) மேற்கோள் வாய்ப்பாடு மற்றும் பிற தேவையான தகவல்களைக் கொண்டிருக்கும். இந்த நிலை மூலமொழி இலக்கண மின்னுருபுகளுக்கு இலக்கு மொழி நிகரன்களை அடையாளம் காண அகராதியைப் பயன்படுத்துகிறது.

5.2.7 வாக்கியநிலை உருவாக்கம்

இந்தச் செயல்பாடு தேவையானால் தொடர்பான்களை முதன்மையாக நிரல் மாற்றியமைக்கும். சில வாக்கியநிலை நோக்குகளும் இந்தப் பகுதியால் கையாளப்படும். எடுத்துக்காட்டாக பெயர், வினை உடன்பாடு போன்றவை இங்கு கையாளப்படும். இந்த மட்டம் தொடர்பகுதி மட்ட ஆய்வாகும். வட்டாரச்சொல் வகுப்புகளைக் கொண்ட தொடர் பகுதிகளுக்கு பொருத்தமான

சொற்கள் (அப்படியிருந்தால்) உருவாக்கப்படும். முதன்மை வினை, துணை வினை உடன்பாடு போன்ற உள் தொடர் பகுதி உடன்பாடு இங்கு காணப்படும்.

5.2.8 வட்டாரச்சொல் குழு பிரிப்பான்

இந்தப் பகுதி ஒரு சொல்லுக்கு விபக்தி மற்றும் பிற தகவல்களைக் கணிக்கும் மற்றும் அவற்றைத் தனிச்சொற்களாகத் திருப்பி அனுப்பும்.

5.2.9 உருபனியல் உருவாக்கம்

ஒரு குறிப்பிட்ட இலக்கண வகைக்காக ஒரு சொல்லை உருவாக்க வேர்ச்சொல்லுடன் பொருத்தமான பின்னொட்டுகள் இணைக்கப்பட வேண்டும். உருவாக்கி மொழியின் உருபனியலுக்குத் தக்கவாறு வேர்ச்சொல்லை திரிபுற செய்யும் மற்றும் இலக்குமொழி சொல் வடிவை வெளியிடும். இவ்வாறு உருவாக்கப்பட்ட சொற்கள் முழு இலக்குமொழி வாக்கியத்தை உருவாக்க இணைக்கப்படும்.

5.2.10 மதிப்பீடு செய்தல்

மதிப்பீடு செய்யும் இயக்க நுட்பம் தனி வாக்கியங்களுக்கும் உரைகளுக்கும் வேண்டி இறுதி பயன்பாட்டாளர்களுக்காக நிறுவப்படும். இது புரிந்து கொள்ளுதல் மற்றும் மொழிபெயர்ப்பு தரம் இவற்றிற்காக மதிப்பீடு செய்யப்படும்.

இரண்டாவது நிலையில் மதிப்பீடு ஒழுங்குமுறையின் வேகத்திற்காகவும் பயன்பாட்டிற்காகவும் செய்யப்படும்.

5.3 முடிவுரை

இயந்திர மொழிபெயர்ப்பு இன்றைய காலகட்டத்தில் முக்கியமான ஒரு நடைமுறை செயல்பாடாக மாறும் நிலையை நோக்கி வளர்ந்து கொண்டிருக்கிறது. இணைய தளத்தில் ஏராளமான தகவல்கள் தமிழிலும் பிறமொழிகளிலும் இருக்கின்றன. குறிப்பாக ஆங்கிலத்தில் பல முக்கியமான தகவல்கள் இணைய் தளத்தில் வலம் வருகின்றன. இத்தகவல்களை தமிழ் மட்டும் அறிந்த நம்மால் புரிந்து கொள்ள இயலாது. இதற்கு வேண்டி உடனடி மொழிபெயர்ப்பு தேவை. இணைய தளத்தில் இருக்கும் ஆங்கிலத் தகவல்களை உடனடியாக இயந்திரத்தினாலேயே தமிழில் மொழிபெயர்ப்பு செய்து கிடைக்கப் பெற்மானால் அது மிகுந்த வரபிரசாதமாகும். அத்தகைய திசை நோக்கித்தான் தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகிறது. தமிழிலிருந்து ஆங்கிலத்திற்கும் ஆங்கிலத்திலிருந்து தமிழுக்கும் தமிழிலிருந்து பிற மொழிகளுக்கும் தமிழ்மொழிக்கும் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகள் நடைபெற்று வருகின்றன. தமிழ்ப் பல்கலைக்கழகம், அண்ணா பல்கலைக்கழகம், சென்னைப் பல்கலைக்கழகம், AUKBC நிறுவனம் தமிழ் இணைய பல்கலைக்கழகம் என்பன தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் ஈடுபட்டுள்ளன. இந்திய மொழிகளின் தொழில்நுட்ப வளர்ச்சி (Technological Development Indian Languages) என்ற திட்டத்தின் கீழ் இந்திய மைய அரசின் கருத்துப் பரிமாற்றம் மற்றும் தகவல் தொழில்நுட்ப அமைச்சின் (Ministry Communication and Information Technology) கீழ்வரும் தகவல் தொழில்நுட்ப துறையால் பொருளாதார உதவிப்பெற்று பல நிறுவனங்களும் பல்கலைக்கழகங்களும் ஆங்கில மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு இந்திய-இந்திய மொழிகளுக்கான இயந்திர மொழிபெயர்ப்பு என்பனவற்றில் முழுமூச்சுடன் ஈடுபட்டுள்ளன. இம்முயற்சிகள் யாவும் இணையத் தளத்திலிருந்து மொழிபெயர்ப்பை இலக்காகக் கொண்டு அமைகிறது. இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளின் பேச்சு மற்றும் எழுத்து விரிதரவுகளை திரட்டும் திட்டத்தை பெற்று செயல்படுத்தி வருகிறது.

முற்றிலும் தானியக்க மொழியியல் பெயர்ப்பு இன்றைய காலகட்டத்தில் இயலாத வகையால் கணிப்பொறி உதவியுடன் மனித மொழிபெயர்ப்பை ஊக்குவிக்கிறது. பொதுவான மொழிக்கு அதாவது பொதுவான மொழி உரைக்கு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவது கடினமாகையால் ஒரு குறிப்பிட்ட பொருண்மைகளை வைத்துக் கொண்டு இயந்திர மொழிபெயர்ப்புச் செய்யும் முயற்சிகள் மேற்கொள்ளப்பட்டு வருகின்றன. ஆங்கில-இந்திய மொழிகள், இந்திய மொழிகள்-இந்திய மொழிகள் மொழிபெயர்ப்பு திட்டத்தில் சுற்றுலா மற்றும் உடல் நலம் (Health) என்பன பொருண்மை களங்களாக ஏற்கப்பட்டு மொழிபெயர்ப்பு செய்யும் முயற்சி மேற்கொள்ளப்பட்டு வருகின்றன.

இயந்திர மொழிபெயர்ப்பு என்பது காலத்தின் கட்டாயமாகும். தமிழுக்கான தமிழில் இயந்திர மொழிபெயர்ப்பிற்கான முயற்சிகளில் நாம் ஈடுபட்டே ஆக வேண்டும். இல்லாவிட்டால் இருமொழிய (Bilingual) (ஆங்கிலம், தமிழ்) நிலையிலிருந்து ஒரு மொழிய நிலைக்கு (Mondingual)மாறிவருகின்ற நம் மாணவ சமுதாயத்தைக் காப்பாற்ற இயலாது போய்விடும். இணைய தளத்தில் கிடைக்கும் அறிவு மூலங்களைத் தமிழர்களாய நாம் பெறாது போய் விடுவோம். இதுபோல தமிழில் கிடைக்கும் அரிய தகவல்கள் பிற மொழியினவர்களுக்குப் போய்ச் சேராது. தகவல் பன்மடங்காகப் பெருகிவரும் இக்கால கட்டத்தில் தகவல் பரிமாற்றத்திற்கு இயந்திர மொழிபெயர்ப்பை விடையாகும்.

தமிழ்ச்சொல்வலை

முனைவர் ச. இராஜேந்திரன்
மொழியியல் துறை
தமிழ்பல்கலைக்கழகம்
தஞ்சாவூர்

யூரோ சொல்வலை

சொல்வலையைப்பற்றிக் கூறுகையில் யூரோ சொல்வலையைப் பற்றிக் கூறாது செல்ல இயலாது. ஏனென்றால் இது தான் முதன் முதலில் நடை முறைப்படுத்தப்பட்டு இணையத்தில் பயன்பாட்டில் உள்ளது. யூரோ சொல்வலையானது சொல் உறவுகளால் தொடர்புபடுத்தப்பட்ட சொற்களின் பன்மொழியச் சொல் வலையைத் (multi lingual wordtnet) தருகின்றது. இது மொழிகளுக்கு இடையிலான தேடலைச் சாத்தியமாக்குகின்றது. சொல்வலையானது பெயர்கள், வினைகள், வினையடைகள், பெயரடைகள் இவை பற்றிய செய்திகளைக் கொண்டிருக்கின்றது. மேலும் இவை ஒருபொருள்பன்மொழியக் குழுமக் (synset) கருத்துருவைச் சுற்றி ஒழுங்கமைக்கப்பட்டுள்ளது. யூரோ வலையின் முக்கிய நோக்கம் என்னவென்றால் சிக்கனமான முறையில் பலமொழிகளுக்கிடையே (English, Dutch, Italian, Spanish, German, French, Estoninan, Czech) மிக விரிவான மற்றும் தரம் வாய்ந்த பன்மொழியத் தரவு மையத்தை (Data base) உருவாக்குவதாகும். யூரோநெட்டினால் உருவாக்கப்பட்ட அறிவுவளம் பல தரப்பட்ட எல்லைக்குட்பட்ட பயன்பாட்டாளாரால் பயன்படுத்தப்படுகின்றது.

அகராதியும் சொல் வலையும்

அகராதியானது சொல்லின் எழுத்துவடிவு, பொருள், பயன்பாடு இவற்றைக் காணப் பொதுவாய்ப் பயன்படுத்தப்படுகின்றது. ஆனால் சொல் வலை ஒரு பொருண்மை அகராதியாகும் (semantic lexicon). சொல்வலை சொற்களுக்கு வர்ணனையும் விளக்கமும் மாதிரி எடுத்துக்காட்டுகளும் தருமும். சொற்களுக்கு இடையே உள்ள பொருண்மை உறவுகளை மிகத்தெளிவாகப் பயன்படுத்தக்க விதத்தில் கூறும். உச்சரிப்பு, சொல்லாக்க உருபனியல் செய்திகள், சொல்மூலம் அல்லது வரலாறு இவற்றைத் தராது.

சொற்களஞ்சியமும் சொல் வலையும்

முதல் சொற்களஞ்சியம்/பொருட்புல அகராதி (Thesaurus) ராஜெஸ்ட் என்பரால் உருவாக்கப்பட்டது. இது பொருண்மை ஒழுங்கமைப்பைக் கொள்கையாகக் கொண்டு உருவாக்கப்பட்டுள்ளது. கருத்துரு அடிப்படையில் அமைக்கப்பட்டது. பயன்பாட்டாளர் தங்கள் மனதிலுள்ள கருத்துருவைக் கொண்டு

சொற்களைக் கண்டுகொள்ள இது உதவும். சொல்வலைகளும் சொற்களஞ்சியம் போல அமையும். இதன் கட்டுமான அலகு ஒருபொருள்பன்மொழியக் குழுமம் ஆகும். ஒருபொருள்பன்மொழியக் குழுமத்தில் ஒரு கருத்துருவை வெளிப்படுத்தும் எல்லாச் சொற்களும் குழுமம் செய்யப்பட்டிருக்கும். சொல்வலையைப் பயன்படுத்துபவர் மனதிலுள்ள கருத்துருவைக் கொண்டு சொற்களைக் கண்டு கொள்ள இயலும். ஒருபொருள்பன்மொழியக் குழுமங்கள் உள்ளடங்கு-உள்ளடக்கு உறவு (honymy-hypernymy), சினை-முதல் உறவு (part-whole or meronymy-holonymy), உட்படுத்து உறவு (entailment) போன்ற உறவுகளால் இணைக்கப்பட்டுள்ளன. சொல்வலை கருத்து நிலையையும் சொல் நிலையையும் தெளிவாகப் பிரிக்கின்றது. இந்த வேறுபாடு சொல் உறவுகளுக்கும் கருத்துக்களுக்கும் உள்ள வேறுபாட்டில் பிரதிபலிக்கின்றது. சொற்களஞ்சியத்திற்கு அப்பாற்பட்டு சொல்வலையில் சொற்களுக்கும் கருத்துருக்களுக்கும் உள்ள உறவுகள் வெளிப்படையாகக் கூறப்பட்டு அடையாளப்படுத்தப்பட்டுள்ளன. பயன்படுத்துபவுபவர் உளவுகளைத் தெரிந்தெடுத்து, ஒரு கருத்துருவிலிருந்து மற்றொரு கருத்ருத்துருவுக்குச் சென்று கருத்துரு வெளியிடத்தில் வலம் வர இயலும்.

சொல் வலையில் உறவுகள்

சொல்வலை கருத்துரு-பொருண்மை உறவுகளுக்கும் சொற்களை இணைக்கின்ற சொல் உறவுகளுக்கும் வேறுபாடு காட்டுகின்றது. ஆனால் சொல் நோக்கில் கவனம் செலுத்துபவர் முதன்மையாகச் சொல் உறவுகளைப் பயன்படுத்துவர். சொல்வலை பொருண்மை உறவுகளால் ஒழுங்கமைக்கப்பட்டுள்ளது. சொல் உறவுகள் ஒருபொருள்பன்மொழியக் குழுமங்களுக்கு இடையே உள்ள காட்டிகளாகும். சொல்வலை வேறுபட்ட தொடரியல் வகைப்பாடுகளில் அடங்கும் சொற்களை இணைக்கும் அடுக்கு உறவுகளைக் கொண்டிருக்கவில்லை. நான்கு முக்கியத் தொடரியல் வகைப்பாடுகளான பெயர், வினை, பெயரடை, வினையடை என்பன தனித்தனியாகத் தரப்பட்டுள்ளன. பெயர்கள் படிநிலை அமைப்புகளாகவும்

வினைகள் வேறுபட்ட உட்படுத்து உறவுகளாலும் (entailment relation) பெயரடைகளும் வினையடைகளும் N-பரிமாண உயர்வெளியிடத்திலும் ஒழுங்கமைக்கப்பட்டுள்ளன.

பெயர்ச்சொல்வலை

பெயர்ச்சொல் வலையில் அடிப்படை பொருண்மை உறவு ஒருபொருள்பன்மொழிய உறவாகும். ஒருபொருள்பன்மொழியக் குழுமங்கள் அடிப்படைக் கட்டுமான அலகுகளாகும். ஒருபொருள்பன்மொழியக் கருத்துச்சாயல் என்பது எல்லாச் சூழல்களிலும் ஒன்றையொன்று இடம்பெயர்த்தல் என்பதை உட்படுத்தாது. இடம்பெயர்த்தல் என்ற அளவியால் இயற்கை மொழிகளுக்குத் குறைந்த எண்ணிக்கையிலான ஒருபொருள்பன்மொழிகள் மட்டுமே வரும். சொல்வலலயில் ஒருபொருள்பன்மொழிகளைக் குறைந்தது சில சூழல்களில் இடம் பெயர்க்கலாம். ஒருபொருள்பன்மொழியக் குழுமத்திற்கு ஒரு பொருள்தான் உண்டு.

எ.கா.

(நூல், புத்தகம், புஸ்தகம்) "படிப்பதற்கு ஏற்றவகையில் அட்டைப்போட்டு இணைத்த அச்சடித்த தாள்களின் தொகுப்பு"

சொல்வலையில் ஒருபொருள்பன்மொழியக் குழுமங்கள் பொருண்மை உறவுகளால் இணைக்கப்பட்டுள்ளன. பெயர்களை ஒழுங்குபடுத்தப் பயன்படுத்தப்படும் மிக முக்கியமான உறவு உள்ளடங்கு உறவாகும் (hyponymy). இந்த உறவுதான் சொற்களைச் சொல்படிநிலை அமைப்பில் ஒழுங்குபடுத்துகின்றது.

சொற்படிநிலை அமைப்பு

உள்ளடங்கு உறவுகள் பற்றிய செய்திகள் மரபு அகராதிகளில் விவரணைகளாகத் தரப்பட்டிருக்கும்.

எ.கா.

குயில் – இனிய குரலுடைய கரிய நிறப்பறவை.

பறவை - இரு கால்களும் அலகுகளும் உடைய, உடலில் இருபக்கங்களிலும் பறப்பதற்கு ஏற்றவகையில் சிறகுகளும் கொண்ட விலங்கினம்.

விலங்கினம் – தானாகவே இயங்கும் உணர்வு உறுப்புகளும் செல்லுலோஸ் இல்லாத செல் சுவர்களும் உள்ள உயிரினம்.

உயிரினம் – உயிர் வாழ்கின்ற ஒன்று.

ஒவ்வொரு உள்ளடக்குச் சொல்லும் மிகப்பொதுவான உள்ளடக்குச் சொல்லுக்கு கொண்டு செல்லும். உள்ளடக்கு உறவைச் சொற்களுக்கு இடையே உள்ள உறவாக உருப்படத்தம் செய்ய இயலாது. உள்ளடக்கு உறவு அகராதிப்படுத்தப்பட்ட கருத்துக்களுக்கு இடையே உள்ள உறவாகும். ஒரு சொற்படிநிலை அமைப்பை உள்ளடக்கு உறவால் இணைக்கப்பட்ட ஒருபொருள்பன்மொழியக் குழுமங்களின் தொடர்ச்சியால் மீட்டுருவாக்கம் செய்யலாம்.

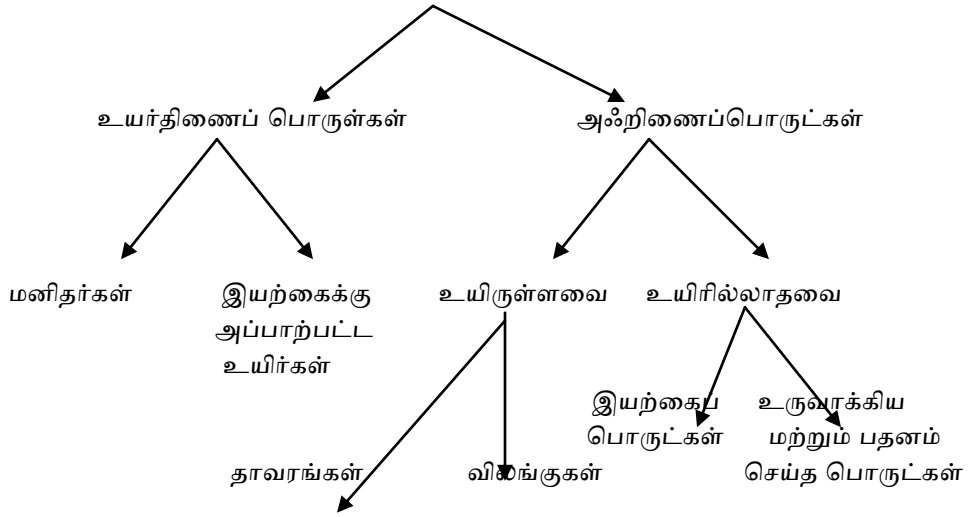
{குயில்}@->{பறவை} @ -> {விலங்கினம்}@ -> {உயிரினம்}

உள்ளடக்கு உறவுக்கு இணையான உள்ளடக்கு உறவையும் தரலாம். உள்ளடங்கு-உள்ளடக்கு உறவுகளால் ஒரு படிநிலை ஒழுங்குமுறை வெளிப்படும். இந்த வகையில் படும் படிநிலைகள் கணியியலார்களால் அறிவை உருப்படுத்தம் (knowledge representation) செய்யும் வழியாகப் பயன்படுத்தப்படுகின்றது.

தனித்தன்மையான தொடக்கிகள்

பெயர்கள் எல்லாவற்றையும் ஒரே படிநிலை அமைப்பில் தரவேண்டி படிநிலைக் கொள்கையை நீட்சி செய்யலாம். சொல்வலை பெயர்களைத் தன்மையான தொடக்கிகள் (unique beginners) அமையப் படிநிலை அமைப்புகளாகப் பகுத்துள்ளது. தனித்தன்மையான தொடக்கிகள் சொற்பொருண்மையியல் பொருண்மைக்கூறாய்வின் பொருண்மைக் கூறுகளுடன் ஓரளவுக்குப் பொருந்தும்.

பருமைப் பொருள்கள்



மூலச்சொல் அமைப்பைக் (ontological structure) காட்டும் வகையிலான சொற்களின் உருப்படுத்தம் சொற்களின் பொருண்மை மரபுரிமைச் செயற்பாங்கை (Lexical inheritance) வெளிக்கொணர்கின்றது.

எ.கா.

{ஊர்தி, வாகனம்}@ -> {நிலத்தில் ஓடும் ஊர்தி}@ -> {பொறியால் ஓடும் ஊர்தி} @ -> நான்குசக்கர ஊர்தி}@ -> {பயண ஊர்தி} @ -> {பேருந்து}

List of 25 unique beginners for noun source files of Euro WordNet

{act, activity}	{natural object}
{animal, fauna}	{natual phenomenon}
{artifact}	{person, human being}
{attribute}	{plant, flora}
{body}	{possession}
{cognition, knowledge}	{process}
{communication}	{quantity, amount}
{event, happening}	{relation}
{feeling, emotion}	{shape}

{food}	{state}
{group, grouping}	{substance}
{location}	{time}
{motivation, motive}	

வேறுபடுத்தும் பொருண்மைக் கூறுகள்

பெயர்களின் படிநிலை அமைப்பு உள்ளடங்கு-உள்ளடக்கு உறவால் உருவாக்கப்பட்டாலும் விளக்கங்கள் ஒரு கருத்துருவை மற்றொரு கருத்துருவிலிருந்து பிரிக்கும் பொருண்மைக் கூறுகளால் தரப்படுகின்றது.

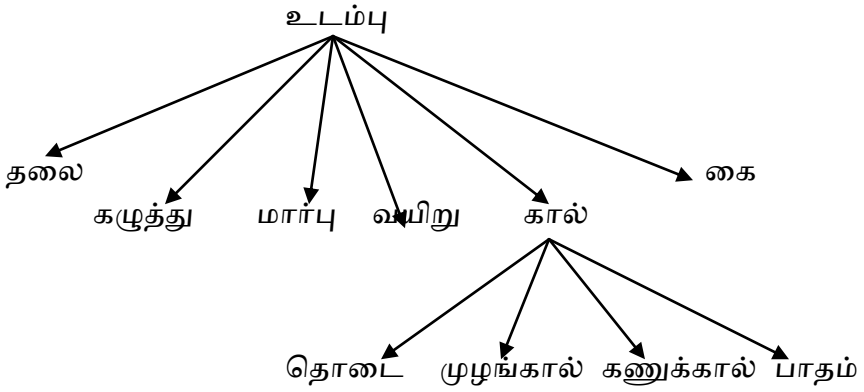
எ.கா.

குயில் – ஒரு பாடுகின்ற சிறிய கரிய நிறப்பறவை
குயில் என்பதை மூன்றுவகைப் பட்ட வேறுபடத்தும் பண்புக்கூறுகளுடன் தொடர்புபடுத்தலாம்.

1. அடைகள்: சிறிய, கரிய
2. பாகங்கள்: அலகு, சிறிய
3. செயல்பாடுகள்: பாடு, பற

சினை-முழு உறவு

சொற்களைப் படிநிலை அமைப்பில் தருவதில் சினை-முழு உறவு முக்கியப் பங்கு வகிக்கின்றது.



சினை-முழுஉறவு பின்வரும் துணை வகைகளை உள்ளடக்கும்:

1. முழுப் பொருளுக்கும் அதன் உறுப்புகளுக்கும் இடையிலான உறவை வெளிப்படுத்துவன (பாகம், எ.கா. கை – விரல்)
2. முழுமைக்கும் அதிலிருந்து பிரிந்த பகுதிக்கும் உள்ள உறவை வெளிப்படுத்துவன (பகுதி, எ.கா. துண்டு – உலோகம்)
3. இடங்களுக்கும் விரிந்த இடங்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (இடம், எ.கா. பாலைவனச்சோலை: பாலைவனம்)
4. குழுமத்திற்கும் அதன் அங்கத்தினர்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (எ.கா. மந்தை - ஆடு)
5. பொருள்களுக்கும் அது உருவாக்கப்பட்ட உருப்பொருள்களுக்கும் உள்ள உறவை வெளிப்படுத்துவன (ஆன, எ.கா. புத்தகம் – காகிதம்)

பின்வரும் அட்டவணையில் பெயர்வலையில் கூறப்பட்டுள்ள பொருண்மை-சொல் உறவுகள் பட்டியலிடப்பட்டுள்ளன:

உறவுகள்	துணை வகைகள்	எடுத்துக்காட்டுகள்
ஒருபொருள் பன்மொழி		புத்தகம், நூல்
உள்ளடங்கு-உள்ளடக்கு		விலங்கு-பலூட்டி
உள்ளடக்கு-உள்ளடங்கு		பசு-பாலூட்டி
முழு-சினை	முழுமை-பாகங்கள்	மேசை-கால்
“	பாகம்-முழுமை	சக்கரம்-வண்டி
“	அங்கத்தினர்கள்-குழு	படைத்தலைவர்-படை
“	குழு-அங்கத்தினர்கள்	துறை-பேராசிரியர்
“	பகுதி-முழுமை	துளி-கண்ணீர்
“	இடம்-பரந்த இடம்	பாலைவனச்சோலை-பாலைவனம்
“	பொருள்-உருப்பொருள்	புத்தகம்-தாள்
ஈரிணை எதிர்நிலை	நிரலாக்கம் செய்யத்தக்கது	நல்லவன் – கெட்டவன்
“	துணை எதிர்நிலை	பகல் – இரவு

“	தனிப்பட்டவை	அஃறிணை– உயர்திணை
“	துருவ எதிரிநிலை	ஆண் – பெண்
“	பரஸ்பர சமூகப்பாத்திரங்கள்	மருத்துவர் – நோயாளி
“	சொந்தங்கள்	அம்மா – மகள்
“	கால உறவுகள்	காலை – மாலை
“	இட உறவுகள:செங்கோண	வடக்கு-கிழக்கு, மேற்கு
“	உறவுகள்	
“	இட உறவுகள்: நேர் எதிர்நிலை	வடக்கு –தெற்கு
“	பல்லிணை எதிர்நிலை: சங்கிலி	ஒன்று, இரண்டு, மூன்று
“	சுற்று	ஞாயிறு, திங்கள், செவ்வாய், புதன், வியாழன், வெள்ளி, சனி

வினைச் சொல் வலை

ஒரு குறிப்பிட்ட செயலை அல்லது தெரிவைக் குறிக்கும் சொல் வினைகள் ஒரு மொழியின் சொல் மற்றும் தொடரியல் வகைப்பாட்டில் முக்கிய பங்கு வகிக்கின்றது. இதன் பயனிலை – பங்கெடுப்பாளர் அமைப்பு (predicate argument structure) இது வரும் வாக்கியத்தின் சாத்தியமான அமைப்புகளைத் தீர்மானிக்கின்றது.

எ.கா.

காற்று வீசுகின்றது.

அவன் சென்னையிலிருந்து தஞ்சாவூர் வந்தான்.

அவள் மாம்பழம் சாப்பிடுகிறாள்.

அவர் அவளிடமிருந்து ஒரு லட்ச ரூபாய்க்கு ஒரு கார் வாங்கினார்.

வினைகளைச் சொற்களங்களாகப் பிரித்தல்

யூரோ சொல்வலை வினைகளை 15 சொற்களங்களாகப் பிரிக்கின்றது. தமிழ்ச் சொல்வலை வினைகளைத் தற்காலிகமாக (இராஜேந்திரன், 2001) பௌதிக வினைகள் (*பெய், வீசு*), உடல்கூறு வினைகள் (*பிரசவி, வியர்*), புலனுணர்வு வினைகள் (*கச, குளிர்*), உணர்ச்சி வினைகள் (*கோபப்படு, வேதனையடை*), அறிவுசார் வினைகள் (*ஊகி, கணி*), கருத்துப்பாற்ற வினைகள் (*பேசு, இகழ்*), சலன வினைகள் (*நகர், நட*), தாக்க வினைகள் (*அடி, உடை*), உடைமைமாற்ற வினைகள் (*கொடு, விடு*), பல்கூட்ட வினைகள் (*சமையல் செய்*) எனப்பிரித்துள்ளது.

வினைகளுக்குத் தனித்தன்மையான தொடக்கிகள்

வினைச் சொற்களஞ்சியத்தைப் பொருண்மைக் களங்களாகப் பிரிப்பது வினைத் தரவுகளை ஒழுங்கு முறை படுத்துவதோடு சொற்களஞ்சியத்திலுள்ள எல்லா வினைகளுக்கும் ஒரு பொதுவான வேர் அல்லது தனித்தன்மையான தொடங்கிகள் இல்லாத குறையையும் நிவர்த்தி செய்கின்றது.

ஒருபொருள்பன்மொழியக்குழுமங்களாக வினைகள்

பெயர்களைப் போல வினைகளையும் ஒருபொருள் பன்மொழியக்குழுமங்களாகக் குழும இயலும். ஆனாலும் நாம் ஒருபொருள்பன்மொழி என்பது ஒன்றையொன்று இடம் பெயர்க்கும் சொற்கள் என்று வரையறை செய்தால் வினையில் ஒருபொருள்பன்மொழிகள் இல்லாமல் போகும் அல்லது மிக அரிதாகத் தான் இருக்கும்.

வினைகளின் பொருண்மைக் கூறாய்வு

வினைகளை அவற்றின் பொருண்மைப் பண்புக்கூறுகளால் விவரணை செய்ய இயலும். வினைகளைச் சிறு பொருண்மைக்கூறுகளாகப் பிரிக்க இயலும் தன்மை காரணமாக எளிய செயல்களால் கலவைத் தன்மையான செயல்கலை விளக்க இயலும்.

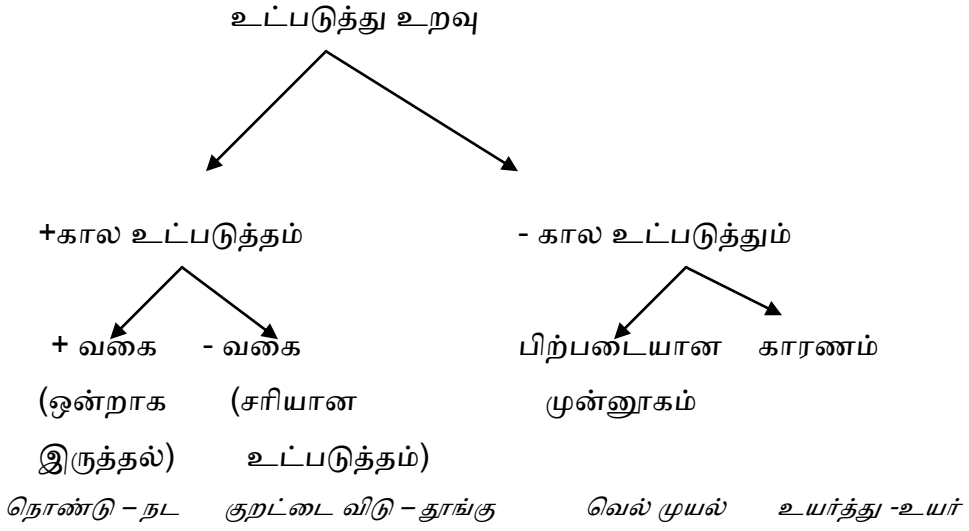
கொல் – சாகச் செய்தல்

எறி - ஒருவர் தன் கையிலிருந்து பொருளை விசையால் போகச் செய்தல்

ஓட்டு - ஒருவர் ஒன்றை ஓடச் செய்தல்

வினைகளுக்கிடையிலான சொல் மற்றும் பொருண்மை உறவுகள்

நாம் நான்கு வகையான உட்படுத்து உறவுகளை (entailment) வெளிப்படுத்தலாம்:



வினைகளுக்கிடையிலான பல்பொருள் ஒருமொழியம்

பெயர்களை விட வினைகளின் எண்ணிக்கை குறைந்தவை. வினைகளின் பல்பொருள் ஒருமொழியம் பெயர்களின் பல்பொருளொருமொழியத்தைக் காட்டிலும் அதிகம். வினைகளின் பொருண்மை நெகழ்ச்சி அதன் பகுத்தாய்வைக் கடினப்படுத்துகின்றது. வினைகள் அவை எடுக்கும் பங்கெடுப்பவர் அமைப்பு அடிப்படையில் (argument structure) தங்கள் பொருளை மாற்றும். ஆனால் பெயர்களின் பொருண்மை வினைகளின் பொருண்மைகளைக் காட்டிலும் நிரந்தரமானது.

தொடரியல் பண்புகளும் பொருண்மை உறவுகளும்

தற்போது அகராதியில் சொற்களுக்கு அதன் தொடரியல் பண்புக்கூறுகளைத் தரும் போக்கு காணப்படுகின்றது. வினைகளைப் பொருண்மை உறவுகளாக மட்டும் பார்ப்பதும் தொடரியல் பற்றிய சில செய்திகளைத் தரும். வினையைப்பற்றி தெரிந்து கொள்ள வினையின் தொடரியல் பண்புகளைச் சொல் வலையில் தருவது குறித்த சாத்தியம் ஆயப்பட வேண்டும்.

உறவுகள்	**விளக்கம்/துணை வகை**	**எடுத்துக்காட்டு**
ஒருபொருள் பன்மொழியம்	இடம்பெயர்க்க இயலும் செயல்கள்	தூங்கு – உறங்கு
பகுதி – முழுமை	உட்படும் – உட்படுத்தம் செயல்	பற – பிரயாணி
வகை உறவு	செயல் – துணை வகை	நட – நொண்டு
உட்படுத்து உறவு	செயல் – காரணச்செயல்	உயர் – உயர்த்து
“	செயல் – முன்னூகச் செயல்	
எதிர்நிலை	எதிர்நிலை	கூடு – குறை
	மறுதலை	வில் – வாங்கு
	திசை எதிர்நிலை	புறப்படு – வந்துசேர்

பெயரடை மற்றும் வினையடைச் சொல் வலை

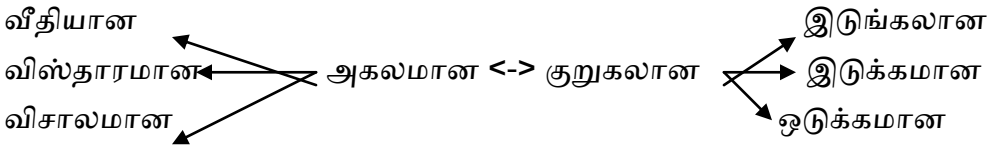
சொல் வலை பெயரடைகளை இரு முக்கியமான வகுப்புகளாகப் பிரிக்கின்றது: வர்ணனைப் பெயரடைகள் (descriptive adjectives) மற்றும் தொடர்புப் பெயரடைகள் (relational adjectives). வர்ணனைப் பெயரடைகள் தலைப் பெயர்களுக்கு இருதுருவ அடைகளின் மதிப்பீட்டைத் (values of bipolar attributes) தருகின்றது; எனவே ஈரிணை எதிர்நிலைகள் (binary oppositions) (எதிர்மறை-antonymay) தருகின்றது; மற்றும் பொருண்மை ஒற்றுமையால் (similarity of meaning) ஒழுங்கமைக்கப்படுகின்றது.

வர்ணனைப் பெயரடை (descriptive adjective): பெரிய, கனமான

தொடர்பப் பெயரடை (relational adjective): பொருளாதார, சகோதர

குறிப்பு அடைசெய்யும் பெயரடை (referemce modifying adjectives): பழைய, முன்னாள்

ஒரு வர்ணனை அடை ஒரு பெயருக்கு ஒரு மதிப்பு தரும். எடுத்துக்காட்டாக, *கனமான, இலேசான* என்பது *எடை* என்பதன் அடைமொழிகளாகும்; *தாழ்ந்த உயர்ந்த* என்பன *உயரம்* என்பதன் அடைமொழிகளாகும். எதிர்மறை (antonymy) தான் வர்ணனைப் பெயரடைகளில் அடிப்படையானதாகும். எதிர்மறை () தான் வர்ணனைப் பெயரடைகளில் அடிப்படையானதாகும். எதிர்மறை ஒருபொருள்பன்மொழியத்தைப் போன்று சொல்வடிவுகளுக்கு இடையிலான உறவாகும். எதிர்மறைப் பெயரடைகள் அடைமொழியின் எதிர்செய்யும் மதிப்புகளை வெளிப்படுத்துகின்றது (எ.கா. *கனமான, இலேசான* என்பன எடை அடைமொழியின் எதிர் துருவங்களாகும்.) நேரடியான எதிர்மறைகள் இல்லாத பெயரடைகள் (எதிர்மறை ஆற்றல் குறைவாய் உள்ள பெயரடைகள்) எதிர்மறைப் பெயரடைகளுடன் "போல" உறவு கொண்டிருக்கும்:



உறவுகள்	**தொடர்புபடுத்தும் சொல்வகைப்பாடு**	**எடுத்துக்காட்டு**
எதிர்மறை (நிரலாக்கம் செய்ய இயலுவது)	பெயரடை/பெயரடை	அழகான/குரூரமான
எதிர்மறை (நிரலாக்கம் செய்ய இயலாதது)	பெயரடை/பெயரடை	உயிருள்ள/செத்த
ஆக்கவடிவு	பெயரடை/பெயர்	அழகான/அழகு
அடைமொழி	பெயர்/பெயரடை	வடிவம்/சின்ன
தொடர்பு	பெயரடை/பெயர்	பொருளாதார/பொருளாதாரம்
போல	பெயரடை/பெயரடை	பாரமான/கனமான

சொல்வலையை திட்டமிட்டு நடைமுறைப்படுத்தல்

சொல்வலையை நான்கு ஒழுங்கமைப்புகளாகப் பகுக்கலாம்.

1. சொல்வள ஒழுங்குமுறை (Lexial resource system)
2. ஒருங்கிணைக்கும் ஒழுங்கு முறை (Compiler system)
3. சேகரிப்பு ஒழுங்கு முறை (Storage system)
4. மீளப்பெறும் ஒழுங்கு முறை (Retrieval system)

அகராதியியலார் சொல்வலையின் மூலக் கோப்புகளை எழுதுகின்றனர். இது சொல்வள ஒழுங்குமுறையை அமைக்கும். ஒருங்கிணைக்கும் ஒழுங்குமுறை சொல்வள கோப்புகளை தரவுமைய வடிவாக உருவாக்குகின்றது. சேகரிப்பு ஒழுங்குமுறை சொல்வள ஒழுங்குமுறைக்கும் ஒருங்கிணைப்பு ஒழுங்குமுறைக்கும் இடையில் செயலாற்றுகின்றது. மீளப்பெறும் ஒழுங்குமுறை பின்வரும் செயல்பாடுகளை உள்ளடக்கும்:

1. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை மின் தட்டச்சு அமைப்பாக மாற்றுகின்றது.
2. வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை தகவல் மீட்பு ஒழுங்கமைப்பாக மாற்றுகின்றது.
3. இணைய மீளப்பெறுதல்: வகைப்பாட்டு மற்றும் வலைப்பின்னல் திட்டத்தை குறிதேடலுடன் இணைப்பது.

முடிவுரை

சொல்வலை ஒரு சொல் தரவு மையம் (Lexical database). இதன் முக்கியமான பண்பு அதன் அர்த்தங்களின் வலைப்பின்னல். சொற்களுக்கு இடையே உள்ள ஒருப்பொருண்மையை அளந்து சொற்பொருள் மயக்கம் தீர்க்க சொல்வலை பயன்படுகின்றது. சொல்வலையைப் பயன்படுத்தி படிநிலை அமைப்பில் தரப்படுள்ள ஒரு சொற்களுக்கு இடையே உள்ள பொருண்மை உறவை ஆய்ந்து அச்சொற்களுக்கு இடையே உள்ள ஒற்றுமையை அளக்கலாம். பொருண்மை மயக்கம் தவிர மீட்புச்செயல்பாடுகளுக்கும் சொல்வலை பயன்படுகின்றது. தகவல் தளங்களிலும் தகவல் மையங்களிலும் இருந்து பயன்படுத்துபவரின் கேள்வியால் தேவையான தகவல்களைப் பெற இயலும். சொல்வலை இதற்குப் பெரிதும் உதவும். இயற்கை மொழி ஆய்வுக்கும்

உருவாக்கத்திற்கும் இது பல வழிகளில் கைகொடுக்கும். சொல்வலையைப் பயன்படுத்தி இயந்திரமொழிபெயர்ப்பு செய்வதற்கான முயற்சிகளும் நடைபெற்று வருகின்றன. சொல்வலையில் உட்படுத்தப்பட்டுள்ள சொல் அறிவு அதைத் திறமையான ஒழுங்குமுறைகள் (expert systems), மொழிபெயர்ப்புத் துணைக்கருவிகள், தேடல் இயந்திரங்கள், கற்றல் ஒழுங்கமைப்புகள் மற்றும் தானியங்கு சுருக்கிகள் (antomatic summarizers) போன்றவற்றிற்குப் பொருத்தமானதாகச் செய்கின்றது. யூரோ சொல்வலை போலவே திராவிடமொழிகளின் சொல்வலை (DraviNet) உருவாக்குவதற்கான முயற்சிகள் தொடங்கப்பட்டுள்ளன. இது நிறைவேறினால் மொழிகளைக் கடந்து தகவல்களைப் பெற இயலும். சென்னையில் இதற்காக ஒரு பணிப்பட்டறை நடத்தப்பெற்று ஆரம்பக்கட்டத் திட்டம் வரையப்பட்டுள்ளது. பல காரணங்களால் இது இன்னும் சாத்தியமாகவில்லை.

* * * * *

துணைநூல்கள்

சுப்பையா பிள்ளை, 2000. இயற்கை மொழி ஆய்வு. உலகத்தமிழாராய்ச்சி நிறுவனம், சென்னை.

செவ்வேள் கபிலன். 1994. கணிப்பொறிவழி தமிழ் வினைகளின் பகுப்பாய்வு. புத்தாக்க மொழியியியல் கழகம், சென்னை

பாஸ்கரன், ச. 2004. தமிழில் கணிப்பொறியியல், கணிப்பொறியில் தமிழ். உமா பதிப்பகம், தஞ்சாவூர்.

Anbarasan, N. 2001. “Tamil Unicode-What do we need?”, paper read in National Seminar on Computational Linguistics and Dravidian Languages. CAS in Linguistics, Annamalai University, Annamalai.

Anandan, P, Rajani Parthasarathy, Geetha, T.V. 2001. “Morphological Generator for Tamil”, in Proceedings of the Tamil Internet 2001 Conference.

Annamalai, E. “Corpora Development in Indian Languages”, in Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

Arulmozhi, P., Sobha, L., Kumara Shanmugam, B. 2004. Parts of Speech Tagger for Tamil, Symposium on Indian Morphology, Phonology & Language Egnineering, Indian Institute of Technology, Kharagpur.

Arulmozhi, P and Sobha, L. 2006. A Hybrid POS tagger for a Relatively Free Word Order Language. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages, pages 79-85.

Arulmozi, S. 1998. Aspects of Inflectional Morphology–A Computational Approach. Ph.D dissertation submitted to University of Hyderabad.

Arulmozi, S and M.C. Kesava Murty. 2006. Verb Sense Disambiguation in Telugu-Tamil MT. In Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages pages 107-111.

Balakrishnan, R. 2002. Morphology and Tamil Computing. Paper read in International Seminar on Tamil Computing, February 27, 28, 2002, Madras University.

Baskaran, S. “User Interface with Computers in Tamil-an Overview [*kaNippoRi uraiyaakam*]”, in Proceedings of the Third Tamil Nadu Science Congress, Pondicherry.

Baskaran, S. “Zip’s Law: A perspective study in Tamil”, in Proceedings of Eighth International Conference of Tamil Studies.

Baskaran, S. Report on Standardization of Tamil Key Board, Submitted to the Committee formed for Standardization of Tamil Key board by Tamilnadu Government.

Baskaran, S. “Tamil Computing”, *tamizh vaLarcci*.

Baskaran, S. “Evolution of Tamil computing”, Souvenir, National Symposium on Current Trends in Computer Applications.

Baskaran, S. “Experiment in Implementation of Finite-State-Machine Parser for Tamil Sentences”, in Computer Society of India Communications, Bombay.

Baskaran, S. “Compute as a Language Research Tools,” in Proceedings of the Second Tamilnadu Congress, Madras.

Baskaran, S. 2002. Word Sense Disambiguation of Tamil. MS dissertation submitted to Anna University, Chennai.

Baskaran, S. 2003. “Word Sense Disambiguation in Tamil”, International Journal of Dravidian Linguistics.

Baskaran, S and Vaidehi, V. 2004. “Collocation Based Word Sense Disambiguation using Clustering for Tamil”, International Journal of Dravidian Linguistics 33.1: 13-28.

Baskaran, S. Vijay-Shanker, K. 2003. “Influence of Morphology in Word Sense Disambiguation for Tamil”, in Rajeev Sangal et al (eds.) Recent Advances in Natural Language Processing. Mysore: Central Institute of Indian Languages.

Brochure on ‘Language Technology Products’ of the Resource Centre for Indian Language Technology Solutions-Tamil, Chennai.

Chellamuthu, K.C. et al. Tamil University Machine Translation System (TUMTS), Thanjavur: Tamil University.

Chellamuthu, K.C. 2002. ‘Russian to Tamil Machine Translation System at Tamil University,’ in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, 74-83.

Chithiraputhiran, H. Tamil Lexical Resource: An electronic Data Base. Paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Deivasundaram, N. and Gopal, A. 2003. ‘Computational Morphology of Tamil’ In B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 406-410.

Dhurai Pandi, 2002. “Morphological Generator and Parsing Engine for Tamil Verb Forms (abstract)”, in Kalyanasundaram (ed) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 59.

Durai Pandi. 2002. “English-Tamil Machine Translation System”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 86.

Durai Pandi, 2002. “Morphophonemic Rules for Tamil Computing”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 98.

Elangovan, A. “Optimisation Techniques in Unicode Tamil Font Development”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 30.

Final Report: Development of Corpora of Texts of Indian Languages in Machine Readable form: Part II (Tamil, Telugu, Kannada, Malayalam), TDIL-Corpora Group, CIIL, Mysore January 1995.

Francis Ekka and Ganesan, M. 1994. “Issues on Standard ISCII Codes and Inscript Keyboard”, Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, New Delhi: BPB Publication.

‘The Unicode Standard: A Technical Introduction’ 8/05/2001.

Ganesan, M. 1994. “Functions of the Morphological Analyser Developed at CIIL, Mysore”, in Harikumar Basi (ed.) Automatic Automatic Translation (seminar proceedings), Thiruvananthapuram: ISDL.

Ganesan, M. 1994. “‘A Scheme for Grammatical Tagging of Corpora of Indian Languages”, in B.B. Rajpurahit (ed.) Technology and Languages, Mysore:CIIL

Ganesan, M. 1996. “Relevance on Indian Grammatical Theories for Automatic Translation: chances and challenges”, paper presented in National Seminar of Dravidian Linguists Association, Dravidian University, Kuppam, June 1996.

Ganesan, M. 2000. “Compilation of Electronic Dictionary for Tamil”, Annamalai University, Annamalai Nagar. www.bhasshaindia.cjb.net/G7ganesan.pdf

Ganesan, M. 2003. “Computational Morphology of Tamil”, in B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, pages 399-405.

Ganesan, M and Francis Ekka. 1994. “Morphological analyzer for Indian Languages”, Agarawal and Pani (eds.) Information Technology Applications in Language, Script and Speech, Newdelhi: BPB Publication.

Hewavitharana, S. 2002. ‘A Two Stage Classification Approach to Tamil Handwriting Recognition’ in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, page 118-124.

Information Technology Department. ‘Information Technology – Standardization of Tamil Key Board and Encoding of Tamil. Glyphs – Recommendations of the Sub-committee on Tamil in Information Technology – Accepted – orders – Issued. (G.O. Ms.NO.17 dated” 13 June 1999)

Jayaram, B.D. “Development of Corpora in Indian Languages”, paper presented in the Seminar in the use of Computers in Indian Languages, held in CIIL, Mysore, August 1992.

Kalyanasundaram, K.A. “Comparison of transliteration schemes and key mapping of Tamil fonts”, http://www.Geocities.com/Athens/5180/translit.html 19/12/2000

Kamakshi, S. 2002. “Preliminaries for Digitizing the Personal Pronouns in English into Tamil (Distribution Sensitive Machine Translation Aid–DSMTA)–A Demo Paper”, in

Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 87-97.

Kamakshi, S. 2004. “Machine Recognition and Translation of ‘~ing’ words in English into Tamil through bilingual Machine Tractable Dictionary”, paper presented in International Conference on Indian Lexicography 28-30 January 2004, CASL, Annamalai University, Annamalainagar.

Kamakshi, S. and Rajendran, S. 2004. Preliminaries to the preparation of a machine aid to translate Linguistics Texts written in English into Tamil. DLA Publications, Thiruvanathapuram.

Kasirao, V. “Impact of Information Technology on Information Management Sevices with Special Reference to Tamil Computing: A Study”, paper read in International Seminar on Tamil Computing, February 27, 28. Madras University.

Krishnamoorthy, V. “OCR Software for Printed Tamil Text”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 99-101.

Kumara Shanmugam, B. 2001. “Parse Representation of Tamil Syntax”,

Kumara Shanmugam, B. 2002. “Machine Translation as related to Tamil”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 84-85.

Kumara Shanmugam, B. 2004. Syntactic Parser for Tamil. M.S. (in NLP) dissertation submitted to Anna University, Chennai.

Lyons, J.1977. Semantics, volume 1, New York: Cambridge University Press.

Michael S. Kapalan. 2002. “Unicode and Tamil”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 1.

Muthu Nedumaran. 2002.”Glyph Choices and Techniques for Building Unicode Based Tamil Fonts”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 31.

Natanasabapathy, S. 2002. “Study of Segmental Duration Tami sounds”, paper presented in International Seminar on Tamil Computing, 27-28 February and 1 March, 2002, University of Madras, Chennai.

Nida, E.A. 1975.a. Compositional Analysis of Meaning: An Introduction to Semantic Structure. The Hague: Mouton.

Nida, E.A. 1975.b. Exploring Semantic Structure. The Hague: Mouton.

Navaneethan, P et al. 2004. “Multilingual Word Processor with Online Noun Translator”, paper read in National Seminar on Cross Language Communication Tools (Machine Translation) for Indian Languages, February 25-27, 2004.

Nayeemulla Khan, A, Suryakanth, V, Gangashetty, & Yegnanarayana, B. 2003. “Syllabic Properties of Thee Indian Languages: Implications for Speech Recognition and Language Identification”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Ponnavaiko, M. 2002. “An Investigation on Unicode Standards for Tamil”, in Tamil Internet 2002: Conference Papers, Cennai: Asian Printers, page 29.

Prasanna Venkatesan, P and Chitrakekha T and Kuppuswami P. 2002. “PONN: A Tamil Operating Environment”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers.

Prathibha, P., Ramakrishana, A.G. 2002. “Web-enabled Speech Synthesizer for Tamil”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, pages 134-140.

Prathibha, P., Ramakrishana, A.G. and Muralishankar, R. 2002. “Thirukural-II – A Text-to-speech Synthesis System”, in Proceedings of Tamil Internet 2002 Conference, Cennai: Asian Printers, 126-133.

Rajan, K and Ganesan, M. 2002. “Tools for Corpus Analysis”, paper presented in International Conference on South Asian Linguistics (ICOSAL-4), 3-5 December, 2002, CASL, Annamalai University, Annamalainagar.

Rajendran, S. 2002. “Preliminaries to the Preparation of wordnet for Tamil”, Language in India2:1, www.languageinindia.com

Rajendran, S. 2003. “Dravidian WordNet: a proposal”, in R.M. Sundaram, et al (eds) Facets of Language, Thanjavur

Rajendran, S. 2003. “Creating Generative Lexicon form Dictionaries: Tamil Experience”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Rajendran, S. 2003. “Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages”, Language in India 3:1, www.languageinindia.com

Rajendran, S. 2006. “Shallow Parsing in Tamil: the state of art”, Language in India 6:7, www.languageinindia.com

Rajendran, S. 2006. “Languge Technology in Tamil”, Language in India 6:8, www.languageinindia.com

Rajendran S, Arulmozi S, Ramesh Kumar S, & Viswanathan S. 2003. “Computational Morpohology of Verbal Complex”, in B. Ramakrishna Reddy (ed.) Word Structure in Dravidian, Kuppam: Dravidian University, 376-398.

Rajendran, s. and Baskaran, S. 2002. “Preparation of Electronic Thesaurus for Tamil”, in Proceedings of the International Conference on Natural Language Processing. Mumbai: NCST.

Rajendran, S. and Kamakshi, S. 2002. “Preliminaries to the preparation of a machine translation aid to translate Linguistics Texts in English into Tamil”, Paper read in UGC-SAP National Seminar ‘On Translation’ 7th-9th March 2002, CAS in Linguistics, Annamalai University.

Ramakrishnan, A.G. “Issues in Standardization for Text to Speech in Tamil”, Tamilnet 2001, Kulalumpur, Malasia.

Ramaswamy, V. 2000. Morphological Generator for Tamil. Unpublished M.Phil. dissertation. University of Hyderabad.

Ranganathan, V. 1997. “A Lexical Phonology Approach to Tamil Words by Computer”, International Journal of Dravidian Linguistics 26:1.57-70.

Shanmugan, C. 2001. “Computer Analysis of Simple Sentences in Tamil”, Paper read in UGC-SAP National Seminar on Computational Linguistics and Dravidian Languages, 22-24 February, 2001, CAS in Linguistics, Annamalai University, Annamalainagar

Shanmugan, C. 2002. ”Grammar and Parser: A Program for Syntactic Parsing in Tamil”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Shanmugan, C. “Minimalist Program for Tamil Parsing”,

Sobha, L and Vijay Sundar Ram. 2006. “Noun Phrase Chunker for Tamil Language”, in Proceedings of the First National Symposium on Modeling and Shallow Paring of Indian Languages, pages 194-198.

Suresh, R.M., Arumugan, S., and Aravanan, K.P. 2000. “Recognition of Handwritten Tamil characters using fuzzy classificatory approach”, in Proceedings of the Tamil Internet 2000 Conference, Singapore.

Vaishnavi Ramaswamy. 2000. A Morphological Generator for Tamil. M.Phil Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi Ramaswamy. 2003. A Morphological Analyzer for Tamil. Ph.D Dissertaion Subbmitted to University of Hyderabad.

Vaishnavi, Ramaswamy. 2003. “Parsing in AMPLE, KIMMO & PERL: Nouns in Tamil”, Paper presented in

Vasudevan, V. 2002. “Character Recognition Techniques–A Demo Program”, International Seminar on Tamil Computing, 27-28 February and March 1, 2002, University of Madras, Chennai.

Vasu Renganathan. 2002. “Interactive Approach to Development of English-Tamil Machine Translation System on Web”, in Kalyansundaram K (ed.) Tamil Internet 2002: Conference Papers, Chennai: Asian Printers, pages 68-73.

Viswanathan, S., Rameshkumar, S. Kumara Shanmugam, B. Arulmozhi. S. & Vijay Shnakar, K.2003. “A Tamil Morphological Analyzer”, in Rajeev Sangal, S.M. Bendre & Udaya Narayana Sigh (eds.), Recent Advances in Natural Language Processing. Mysore: CIIL

Winston Cruz, S. 2002. Parsing and Generation of Tamil Verbs in GSMorph. M.Phil. dissertation submitted to the University of Hyderabad.

Yagnanarayana, B. et al. 1992. “Text –to-speech system for Indian Languages”, presented at the Workshop on Computer Applications in Indian Languages held at CIIL, Mysore, August 19-21

Yegnanarayanan B, Rajendran, S, Ramachandaran V.R., Madhukumar, A.S. “Significance of Knowledge Sources for a Text-to-speech System for Indian Languages”, in Sadhana, Academic Proceedings in Engineering Sciences of India.

Yegnanarayanan, B. 1994. “Speech synthesis by Machine”, in B.B. Rajpurohit (ed.) Technology and Languages. Mysore:CIIL.

துணை நூல்கள்

இராஜேந்திரன், ச., 2001 *தற்காலத் தமிழ்ச்சொற்களஞ்சியம்*, தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்.

" 1991 'தற்காலத் தமிழ்ச் சொற்களஞ்சியம் (பாகம் 1),' தமிழ்க்கலை, 9.1. -4:55-76.

" 1999 *'பொருட்புல வகைப்பாடும் சொற்களஞ்சியமும்,'* புலமை, 25.2:47-66.

" 2002 '*இணையம் வழி அயல்நாடுகளில் தமிழ் கற்றலை வளர்த்தல்*, 'தமிழ் ஆய்வில் இன்றைய போக்குகள், 75-81, திருமதி செண்பகம் சுப்பையா அறக்கட்டளை, தஞ்சாவூர்.

இராஜேந்திரன், ச., 2004 *தமிழ்மின்சொற்களஞ்சியம்*

பாஸ்கரன், ச. 2004 *(ஆய்வேடு),* தமிழ்ப்பல்கலைக்கழகம், தஞ்சாவூர்.

Kamakshi, S., & Rajendran, S., 2004 "'Preliminaries to the Preparation of a Machine Aid to Translate Linguistic Texts in English into Tamil', Dravidian Linguistic Association, Thiruvananthapuram.

Rajendran, S., 1995 ' A Machine Bridge for Tamil' (To be published in Proceedings of English World Tamil Conference)

" 1997 'Grammatical Formalism and Computational Analysis of Nominal Compounds in Tamil', South Asian Language Review, Vol 7, No.1., 27-46.

" 2003 'Prerequisite for the Preparation of an Electronic Thesaurus for a Text Processor in Indian Languages', Languages in India 3:1, www.languageinindia.com.

" 2003 'Dravidian WordNet: A proposal,' In R.M. Sundaram et al (edited), 467-497, Facets of Language, Thanjavur.

" 2003 'Creating Generative Lexicon from MRDs: Tamil Experience', In: Rajeev Sangal et al (edited), 83-91, Recent Advances in Natural Language Processing, Proceedings of the International Conference of Natural

Language Processing (ICON 2003),

Rajendran, S., et al 2000 'Tamil WordNet' Proceedings of the First Global WordNet Conference, CIIL, Mysore.

" 2003 'Computational Morphology of Verbal Complex', In:B Ramakrishna Reddy (edited), Word Structure in Dravidian, University, Kuppam & Language in India, 3:4, www.languageinindia.com. April, 2003.

♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠ ♠